

**Adachi, K.**

CS-Tuesday-am-s37

Osaka University, Osaka, Japan (k-adachi@lt.ritsumei.ac.jp)

### Joint Procrustes Analysis in the S-O-R Framework

Let  $\mathbf{X}$  be an  $I$ -stimuli  $\times$   $J$ -responses containing the magnitudes of responses which an organization shows against stimuli. For finding the  $p$  components intermediating between the stimuli and responses, rank- $p$  approximation  $\mathbf{X} \cong \mathbf{A}\mathbf{B}' = \mathbf{Y}\mathbf{S}(\mathbf{Z}\mathbf{S}'^{-1})'$  may be useful with  $\mathbf{S}$  a nonsingular matrix. To identify  $\mathbf{S}$ , I propose three types of joint Procrustes analysis (JPA). The first one is to minimize  $LS_{2\text{-way}} = I^{-1}\|\mathbf{Z}\mathbf{S} - \mathbf{G}\|^2 + J^{-1}\|\mathbf{Y}\mathbf{S}'^{-1} - \mathbf{H}\|^2$  over  $\mathbf{S}$  for given targets  $\mathbf{G}$  and  $\mathbf{H}$ . This is attained by an ALS algorithm in which  $\mathbf{S}$  is reparameterized by its singular value decomposition. The second is to minimize  $LS_{2\text{-way}}$  over  $\mathbf{S}$ ,  $\mathbf{G}$ , and  $\mathbf{H}$ , in order to give simple structures to both  $\mathbf{A} = \mathbf{Z}\mathbf{S}$  and  $\mathbf{B} = \mathbf{Z}\mathbf{S}'^{-1}$ , where some elements of  $\mathbf{G}$  and  $\mathbf{H}$  are constrained to be zeros. This minimization can be attained using the JPA algorithm with a variant of Kiers' (1992) simplimax procedure. In the third approach, 3-way cases are considered where participant numbers  $k = 1, \dots, K$  are attached to symbols as  $\mathbf{X}_k \cong \mathbf{A}_k\mathbf{B}_k' = \mathbf{Y}_k\mathbf{S}_k(\mathbf{A}_k\mathbf{Z}_k'^{-1})'$  and  $LS_{3\text{-way}} = I^{-1}\sum_k\|\mathbf{Y}_k\mathbf{S}_k - \mathbf{A}\|^2 + J^{-1}\sum_k\|\mathbf{Z}_k\mathbf{S}_k' - \mathbf{B}\|^2$  is minimized over  $\mathbf{S}_k$ 's,  $\mathbf{A}$ , and  $\mathbf{B}$ .

**Ahmed, U.S., and Chang, H.-H.**

TS-Monday-am-s11

University of Illinois at Urbana-Champaign, Urbana, IL (usahmed2@uiuc.edu)

### The Impact of Item Selection Method in CAT-DIF Analysis

This paper presents a comparison study among three item-selection methods in computerized adaptive testing (CAT) in terms of their effects on differential item functioning (DIF) analyses. The studied methods are Maximum Information (MI),  $a$ -Stratified (STR), and Constraints Weighted Information (CWI). It is our belief that different item-selection algorithms yield different performance in CAT-DIF. MI has some practical problems such as lower usage of item bank and high item exposure rates for high- $a$ -parameter items. It is important to conduct an empirical study to investigate the performance of some popular CAT-DIF procedures across several promising item selection methods. We propose to use two CAT-DIF methods: Modified-Mantel-Haenszel and CATSIB. Simulation study is conducted to investigate the effect of item-selection methods on the performance of both MMH- and CATSIB-based DIF analyses. In the simulation study, different sample size, impact level, and amount of DIF are investigated. It is expected that the quality of DIF-detection procedures can be enhanced with respect to both Type I error rates and the power level with proper item-selection strategies. The results may help practitioners design or refine their CAT systems in terms of both estimation efficiency and test fairness.

**Bahn, G., and Bryant, F.**

Poster-Tuesday-pm-s50

Loyola University Chicago, Wheaton, IL (gbahn@luc.edu)

### Measuring Educational Values of Korean Americans Based on Confucian Five Moral Codes

The goal of this study was to measure the educational values of first generation Korean Americans by focusing on five Confucian moral codes that have persisted as the fundamental Korean educational philosophy for hundreds of years (Chu, 1993). Survey questions were developed to assess the thoughts, attitudes, beliefs, and values of participants based on the five moral codes. Confirmatory factor analysis (CFA) was used to test three alternative measurement models, the best-fitting of which consisted of the five moral codes. The participants' age and length of U.S. residency against the five moral codes were analyzed through Path Analysis (PA). Results suggest that first generation Korean Americans endorse the values underlying the five moral codes even after living in the U.S. for an average of 16 years. Three of the five codes were significantly correlated with age, suggesting that older parents desire less influence over their child but also are more strongly motivated to support their child and preserve the traditional role of family with husband as provider and wife as housekeeper. Additional results indicate that the longer Korean parents have lived in America, the less they endorse the Confucian codes, except for supporting their child's education financially.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Barrada, J.R.**, Olea, J., Ponsoda, V., and Abad, F.J.  
Universidad Autonoma de Barcelona, Bellaterra, Spain  
(juanramon.barrada@uab.es)

TS-Monday-am-s11

### A Method for the Comparison of Item Selection Rules in Computerized Adaptive Testing

In a typical study of the relative efficiency of two competing item selection rules in computerized adaptive testing, the common result is that they simultaneously differ in accuracy and security, making it difficult to reach a conclusion on which is the most appropriate. This study proposes a strategy to conduct a global comparison of two or more selection rules. A plot showing the performance of each selection rule for several maximum exposure rates is obtained and the whole plot is compared with other rule plots. The strategy has been applied in a simulation study for the comparison of 6 exposure control methods: point Fisher information, Fisher information weighted by likelihood, Kullback-Leibler weighted by likelihood, maximum information stratification method with blocking, progressive method and proportional method. There is no optimal rule for any overlap value or RMSE. The fact that a rule, for a given level of overlap, has lower RMSE than another does not imply that this pattern holds for another overlap rate. A fair comparison of the rules requires extensive manipulation of the maximum exposure rates. The best methods were Kullback-Leibler weighted by likelihood, proportional method and maximum information stratification method with blocking.

**Béguin, A.**  
Cito, Arnhem, The Netherlands (anton.beguin@cito.nl)

IS-Tuesday-am-s32

### Mixed IRT Linking: Combining High-Stakes Tests with a Low-Stakes Anchor.

The quality of the data providing the link between the tests forms that are linked is essential for the accuracy of the linking. In various linking designs it is assumed that common-items or anchor items are proportionally representative of the total test forms in content and statistical characteristics (Cook & Petersen, 1987). The realism of this assumption does not only depend on the content of the common items but also on the condition in which they are administered. Especially responses on items in an external anchor could be influenced by the condition of the administration (e.g., lack of motivation) and consequently be a source of bias in the linking and equating procedures. Using a mixed IRT model (Rost, 1990, 1997; see also Von Davier & Carstensen, 2007) an IRT equating procedure is introduced that is more robust against different response behaviour. Students with different response behaviour on the common items are identified and the link between the forms is based on the responses of students that behave the same on unique and common items. A limited simulation study is conducted to compare performance of IRT equating based on concurrent estimation and IRT equating based on mixed IRT.

**Blozis, S.A.**  
University of California, Davis, CA (sablozis@ucdavis.edu)

CS-Tuesday-pm-s45

### Partially Nonlinear Crossed Random-Effects Models for Longitudinal Data

Longitudinal data are often hierarchically structured in which repeated measures are nested within individuals, with individuals possibly nested within random groups, such as schools or clinics. In these cases, a three-level mixed-effects model may be appropriate. In a different kind of setting, the nesting of individuals may be cross-classified into multiple sampling units. This type of data structure may be handled by a crossed random-effects model. This paper proposes a partially nonlinear, crossed random-effects model for longitudinal data. Repeated measures of cortisol nested within individuals who are nested within both assays and twin pairs are studied. Individuals are treated as indistinguishable within twin pairs. A two-phase, partially nonlinear mixed-effects model is developed to characterize hourly responses over three days.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Bollen, K.A.**, and Bauldry, S.  
University of North Carolina, Chapel Hill, NC (bollen@unc.edu)

CS-Wednesday-am-s65

### Model Identification and Computer Algebra

Multiequation models that contain observed variables or observed and latent variables are common in the social sciences. In order to determine whether unique parameter values exist for such models, one needs to assess the identification of the model. Previously analysts have generally relied on proven rules for determining identification and tests in SEM software packages that determine local identification. In this paper we outline how to use computer algebra systems to determine the global identification of multiequation models with or without latent variables. In the process, we demonstrate how to obtain explicit algebraic solutions for each of the model parameters. We illustrate the procedures through four examples.

**Bolt, D.M.**<sup>(1)</sup>, and Johnson, T.R.<sup>(2)</sup>  
<sup>(1)</sup> University of Wisconsin, Madison, WI (dmbolt@wisc.edu)  
<sup>(2)</sup> University of Idaho

CS-Tuesday-am-s28

### Applications of a MIRT Model to Self-Report Measures: Addressing Score Bias and DIF Due to Individual Differences in Response Style

A multidimensional item response model that accounts for response style factors is considered. The model can be viewed as a multidimensional extension of Bock's (1972) nominal response model and is shown to allow for the study and control of response style effects so as to estimate and ultimately reduce bias in the estimated level of the intended-to-be-measured trait. In the current application, the model is also used to investigate extreme response style as an underlying cause of differential item functioning (DIF) between respondent groups distinguished by education level. The approach is illustrated using the item responses of cigarette smokers to the Wisconsin Inventory of Smoking Dependence Motives (WISDM-68), a self-report measure of tobacco dependence.

**Borsboom, D.**<sup>(1)</sup>, Wicherts, J.<sup>(1)</sup>, and Romeijn, J.-W.<sup>(2)</sup>  
<sup>(1)</sup> University of Amsterdam, Amsterdam, The Netherlands (d.borsboom@uva.nl)  
<sup>(2)</sup> Groningen University, Groningen, The Netherlands

CS-Tuesday-pm-s46

### Measurement Invariance Versus Selection Invariance: Is Fair Selection Possible?

This paper shows that measurement invariance (defined in terms of an invariant measurement model in different groups) is generally inconsistent with selection invariance (defined in terms of equal sensitivity and specificity across groups). In particular, when a unidimensional measurement instrument is used, and group differences are present in the location but not in the variance of the latent distribution, sensitivity and positive predictive value will be higher in the group located at the higher end of the latent dimension, whereas specificity and negative predictive value will be higher in the group located at the lower end of the latent dimension. When latent variances are unequal, the differences in these quantities depend on the size of group differences in variances, relative to the size of group differences in means. The effect is shown to originate as a special case of Simpson's paradox, which arises because the observed score distribution is collapsed into an accept/reject dichotomy. Possible methodological solutions to the problem are discussed.

**Bouwmeester, S.**, van Rijen, S., and Sijtsma, K.  
Erasmus University, Rotterdam, The Netherlands (bouwmeester@fsw.eur.nl)

CS-Tuesday-am-s28

### Constructing a Scale to Measure Phonological Awareness

## ABSTRACTS OF THE CONTRIBUTIONS

---

Segmentation is the ability to cut words into phonemes. Numbers of studies have shown that the ability to segment words into phonemes is an important predictor of early reading. However, due to the fact that researchers used different sets of words that differ in difficulty level to measure segmentation performance, the relationship between segmentation ability and reading varies between studies and is hard to interpret. Previous research showed that properties of the words influence the difficulty to segment words. The aim of this study was to construct a scale that enabled measurement of segmentation ability independent of the kind and difficulty of the words that have to be segmented. An itemset of 45 pseudo-words was administered to 597 Dutch children from kindergarten to grade 2. Models from item response theory were used to investigate the dimensionality and scalability of the itemset. Moreover, the LLTM was used to investigate whether the item properties could predict the item difficulties. We concluded that the segmentation performance could well be explained by one ability. The Rasch model had a reasonable fit on most items. The fit of the LLTM model indicated that the item properties can predict the difficulty level to segment words well.

**Bovaird, J.A.**, Chumney, F., and Wu, C.

[CS-Wednesday-am-s64](#)

University of Nebraska-Lincoln, Lincoln, NE (jbovaird2@unl.edu)

### On the Finite Population Correction in Multilevel Modeling: Implications for Nesting Within Geographic Region

Computation of traditional standard errors assumes that the obtained sample is constructed through random sampling with replacement from an infinitely large population or a practically infinite, yet finite, population. However, most sampling is conducted without replacement from finite populations. In some complex sampling applications in psychology and education it is possible to obtain proportionally large samples or near-census sampling at the macro-levels, especially when sampling from finite geographical locations as in educational testing, cross-cultural research, and behavioral ecosystems modeling. Yet, typical hypothesis testing at the macro-levels utilizes traditional standard errors. The finite population correction (fpc) quantifies the degree of increased precision achieved when the sampling process approaches census sampling by down-weighting the traditional standard error inversely proportional to the difference in size between the obtained sample and the intended population. Guidelines from survey statistics suggest that the fpc may be effective when sampling as little as 5% of the population. This paper evaluates the impact of the fpc on testing macro-level predictors in a multilevel structural equation model of school readiness with students nested within county from a Midwestern state. A small simulation study is included to evaluate published guidelines for the sample size necessary to realize fpc effectiveness.

**Braeken, J.**, Tuerlinckx, F., and De Boeck, P.

[CS-Wednesday-am-s58](#)

K.U. Leuven, Belgium (johan.braeken@psy.kuleuven.be)

### Model Selection in Copula IRT Models for Local Item Dependencies

One of the key assumptions in modern measurement models of item response theory is conditional independence. Local item dependencies (LID), which are violations of conditional independence within specific item subsets, can be accounted for by means of the recently introduced copula IRT models in which this additional association unmodeled by the latent trait, is captured by a copula function. The nature of the LID, whether it is stable over the latent trait or in/decreases dimensionally, is determined by the specific choice of copula. The current study investigates the feasibility of identifying different LID association structures within these copula IRT models. This essentially comes down to a model selection issue: How to properly offset badness-of-fit by model complexity to prevent poor generalization and make a choice between these functionally different and non-nested copula IRT models. Both traditional as well as relatively unexplored approaches such as Minimum Description Length (MDL) and model weights through continuous model expansion are implemented in our search for a proper way of tackling this problem.

**Brasfield, J.**, Roxbury, T., and Ackerman, T.

[CS-Tuesday-am-s36](#)

UNC-Greensboro, Greensboro, NC (jonbrasfield@gmail.com)

## ABSTRACTS OF THE CONTRIBUTIONS

---

### A Simulation Study to Detect Dimensionality Across Subgroups of Examinees

This project uses simulated and real response data to examine the conditions under which differences in IRT dimensionality can be detected for subgroups of examinees. This study is intended to emulate the testing of elementary students using a math test with word problems. For high reading ability students, the data may be essentially unidimensional (math ability) and for low reading ability students the data may be two-dimensional (reading and math). This project involves the generation of both raw reading scores and math response data, based on distributions of reading and math abilities. Low reading ability students will have math responses generated using their math and reading abilities and high reading students will use only math ability. Data will be generated under various conditions related to the difference in reading ability between the two groups, the discrimination of the word problems, and the ratio of word problems to pure math problems. These data will then be analyzed by taking the highest and lowest reading scores and assessing the dimensionality of those students' math scores. Parallel analyses will be conducted using real data. This research aims to provide evidence that teachers need to consider the role literacy may play within a classroom mathematics test.

**Brossman, B.**, and Lee, W.-C.

CS-Wednesday-am-s64

University of Iowa, Iowa City, IA (bradley-brossman@uiowa.edu)

### A Comparison of Confidence Intervals and Tolerance Intervals for Stratified Domains under the Compound Binomial Model

Standard errors of measurement vary across true score levels. In attempt to adequately estimate standard errors as a function of true scores, Lord (1955, 1957) derived a mathematical model, based on the binomial distribution, for dichotomously scored items. This model was used to estimate conditional standard errors for “unstratified” (or single) content domains. Feldt (1984) extended Lord’s derivation to estimate conditional standard errors for scores summed across stratified domains. This derivation was based on the compound binomial distribution. Often, these error estimates are used to construct confidence intervals for which an examinee’s true score may be likely. Depending on the nature of the statistical assumptions behind such intervals, often these interval procedures do not contain the anticipated  $\gamma\%$  coverage. The purpose of the present paper is to 1) create Wald’s confidence intervals, score confidence intervals, and tolerance intervals for stratified domains and 2) empirically determine, based on a series of simulation studies, which of the intervals provide the most accurate coverage according to specified  $\gamma\%$  levels. Although previous authors have investigated both theoretical and empirical properties of various intervals, none have investigated coverage accuracy under the compound binomial model across stratified domains.

**Browne, M.W.**, and Liang, L.

CS-Wednesday-am-s65

The Ohio State University, Columbus, OH (browne.4@osu.edu)

### Locating Person Points on the Circumplex

A data model that generates the Circumplex correlation structure is considered. In this model persons are represented by points that are located on the circumference of the same circle that represents the variable points. A person’s score on a particular variable is related to the distance of the Person point to the variable point along the circumference of the circle. A method for locating person points on the Circumplex is described. Person profiles are simultaneously fitted to the corresponding circumplex data. Examples are given.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Cai, L.**

IS-Wednesday-am-s52

University of North Carolina, Chapel Hill, NC (cai@unc.edu)

### A Metropolis-Hastings Robbins-Monro Algorithm for Maximum Likelihood Estimation in Latent Variable Models

The Robbins-Monro (RM) algorithm is well known in the engineering and optimization literature. Since its inception, it has been primarily used for the purpose of identification and adaptive control of dynamical systems. Recent progress in the use of Markov chain Monte Carlo techniques for the estimation of nonlinear mixed models has led to a dominating insight that the RM method can be combined with MCMC to arrive at a new algorithm – Metropolis-Hastings Robbins-Monro (MH-RM) – for maximum likelihood estimation in latent variable models that is as flexible as MCMC and still preserving the point-wise convergence of RM. Some theoretical properties of the MH-RM algorithm will be discussed, and real and simulated data will be used to illustrate the potential advantage that MH-RM holds against existing methods in fitting complex latent variable models, including dynamical systems models.

**Cardinale, J., Verkuilen, J., and Jeltova, I.**

Poster-Tuesday-pm-s50

City University of New York, Whitestone, NY (jcardinale@gc.cuny.edu)

### Regression in Cross-Cultural Attitude Data that Features Non-Constant Variances

Heterogeneity is observed in cross-culture data when survey respondents treat attitude scales differently depending on their cultural background. For example, Asian cultures have been observed to favor less extreme values across attitude scales, where Mediterranean cultures are observed to choose the extremes of these scales. In addition, there are often different response styles. In other words, different points on an attitude scale may have different meanings depending on the culture. (Dolnicar, 2006). Due to these problems, fixed and random error variances are often non-constant, or heteroscedastic, across all levels of a statistical model. Modern statistical forms of general linear models, particularly mixed linear models, can overcome both fixed and random heteroscedacity. By properly modeling covariance structures and random intercept terms, mixed linear models can properly weight segments of data that are highly variable. This avoids the problems of the traditional z-scoring method because the mixed model shrinks the estimated mean outcome toward the overall mean, and therefore reduces the mean-squared error. A simulation study is developed to compare heteroscedasticity models by examining the accuracy of estimated covariance matrices. It is shown that the mixed linear model produces more accurate covariance estimates, and therefore, more accurate coefficients than traditional regression. This method is related to the problem of differential item functioning (DIF) in item response theory (IRT). For illustration purposes, a presentation with empirical data will be provided.

**Carstensen, C.H., and Wilson, M.**

IS-Tuesday-am-s34

IPN University of Kiel, Kiel, Germany (carstensen@ipn.uni-kiel.de)

### An Application of IRT Models to Causal Inferences from a Pre- and Posttest Design

In this paper, an item response model for analyzing the changes in student achievements from a large scale assessment over two time points is discussed. The motivation for this model is a longitudinal study conducted within the PISA assessments in Germany where the same students were assessed in grades nine and ten. The central research questions relate to the analysis of learning gains, and, especially, the impact of predictors such as socio economic status on those gains. The analyses model includes a measurement part using a generalized Rasch model assuming both time points measure the same ability. In the structural part of the model the assessments from both time points are modeled two-dimensionally, and the effects of predictors are estimated in a latent regression. The model is estimated using ConQuest. With the use of multiple imputations (Rubin, Mislevy et al.), regression models of the grade ten achievement on the grade nine achievement and some predictors give unbiased estimates. Results from a simulation study that support this finding will be presented.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Casabianca, J.,** and Lewis, C.  
Fordham University, Bronx, NY (Casabianca@fordham.edu)

TS-Tuesday-am-s38

### Equivalence Testing for DIF

To assess Differential Item Functioning (DIF) using the Mantel-Haenszel (MH) procedure, the null hypothesis of no DIF is tested. However, failing to reject the null hypothesis is *not* the same as establishing no DIF. Fairness can be better achieved by demonstrating a certain level of equivalence instead of failing to detect a difference. In the current study we apply statistical equivalence testing to DIF analysis, hereafter denoted as *Equivalent Item Functioning* (EIF). A test for EIF uses two one-sided tests to include two intervals of values to define “difference” as the null hypothesis and an alternative hypothesis to include an interval of values to define “equivalence.” In a Monte Carlo simulation we use the Rasch model to generate 19 item responses for two groups of examinees. In addition to both the standard null hypothesis test for the MH estimator and equivalence tests, a confidence interval approach to testing equivalence is performed. Factors varied are: sample size, inclusion or exclusion of the studied item in the criterion, and level of DIF in the studied item. Type I and type II error rates are compared to evaluate both methods and to see how equivalence testing performs when excluding the studied item.

**Cella, D.**  
Evanston Northwestern Healthcare and Northwestern  
University Feinberg School of Medicine, Evanston, IL (d-cella@northwestern.edu)

SA-Monday-am-s4

### IRT Modeling Applied to Self-reported Health and Quality of Life: The Patient Reported Outcomes Measurement Information System (PROMIS)

PROMIS is a publicly-funded research group of investigators from academic institutions and the National Institutes of Health (<http://www.nihpromis.org>). From 2004-2008, we developed, refined and tested nearly 1,000 self-report questions about physical, mental and social health. We administered these questions on an electronic (internet) platform, to a cross sectional sample of approximately 20,000 people from the general US population and selected clinical samples. Using a combination of classical methods to test dimensionality and item response theory (IRT) modeling, we derived nine (9) calibrated item banks that measure unidimensional concepts of fatigue, pain impact, pain behavior, physical function, depression, anxiety, anger, satisfaction with participation in social roles, and satisfaction with participation in discretionary social activities. We also developed and tested item banks in parallel domains for pediatrics, as well as adult banks of sleep/wake disturbance and cancer-specific issues. This presentation will review item bank development and testing, and compare the precision of these item banks and their derivative tools (short forms and computerized adaptive testing; CAT) with existing "legacy" instruments measuring the same concepts. It will illustrate that the precision of CAT and PROMIS short forms outperforms the "legacy standards." Opportunities for collaboration with PROMIS investigators going forward will also be discussed.

**Chajewski, M.,** and Lewis, C.  
Fordham University, Bronx, NY (chajewski@fordham.edu)

CS-Monday-am-s10

### Impact of Missing Data Imputation on Effect Size Estimation in Categorical Data

Presently, little research has been conducted regarding the impact of missing data (and its subsequent imputation) on the estimation of effect sizes in categorical data analysis. This research provides a comprehensive analysis of such impact and aims to create concise guidelines in dealing with missing observations when estimating effect sizes. An empirical simulation was conducted based on a sample of juveniles released from a correctional center in Virginia (N = 402) a comparison of juveniles' Substance Abuse Screening Inventory scores (SASSI) and their race (White or Non-White). The research design consisted of 12 conditions, wherein Cramer's Vs (as the most widely used measure of strength of association between two nominal variables) was estimated for each condition. For both data missing at

## ABSTRACTS OF THE CONTRIBUTIONS

---

random and data not missing at random, two methods of missing data imputation were used to impute values in datasets with 3%, 15% and 45% of the observations missing. Respective effect sizes together with bootstrap derived standard errors and 95% confidence intervals were compared in order to determine the impact of imputation procedures.

**Chang, S.-W.**

Poster-Tuesday-pm-s50

National Taiwan Normal University, Taipei, Taiwan (shwchang@ntnu.edu.tw)

Effects of Raw-to-Scale Score Conversions on Equating: The Choice of Score Scale

This study investigated the effects of various raw-to-scale score transformation methods on equating when the equipercentile equating method was employed under the random groups design. The four procedures of the linear, normalizing, arcsine, and log-odds transformations were compared. A sample size of 20,000 examinees was used to simulate the data resembling the five tests of the BCTEST assessment administered in 2006 in Taiwan. The effects of the various methods were studied on the equating accuracy as well as the properties of the resulting scale score equivalents based on the criteria of the raw-to-raw and the raw-to-scale score conversions, the central moments of both the Form I scale scores and the scale score equivalents for Form II, the impact on the highest scale scores, and the gap sizes in the conversions for Form II as well. The findings indicated that the scale score equivalents for Form II did not differ substantially from the initial Form I scale scores after equating. The relative behaviors of the various transformation methods remained fairly similar after the equipercentile equating process. Results from this study have helped psychometric researchers and test practitioners gain more insights into the scaling and equating issues while establishing the score scales.

**Chen, P.H.**

Poster-Tuesday-pm-s50

National Taiwan Normal University, Taipei, Taiwan (chenph@ntnu.edu.tw)

The Latent Trait Estimation for the Outliers in Computerized Adaptive Testing Using the Weighted Maximum a Posteriori Estimation

The goal of the study is to develop a method to resolve the regression bias of the outliers in computerized adaptive testing (CAT) when using the maximum a posteriori (MAP) estimation. By adding a weighted coefficient to the prior distribution for each person and changing the coefficient after every item selection stage of CAT, the author develops a weighted maximum a posteriori (WMAP) estimation method. Items responses data had been simulated using the multidimensional random coefficients multinomial logit (MRCML) model, a multidimensional item response model. Ten to fifty items of CAT had been carried out using the traditional MAP and the WMAP estimation. The conditional bias and the root mean square of error (RMSE) had been analyzed. Results indicated that the WMAP estimation resulted in less regression bias than the traditional MAP estimation. The WMAP estimation resulted in lower RMSE for the outliers but higher RMSE for the normal simulees than the traditional MAP. The applications of the weighted maximum a posteriori estimation had been addressed in the study.

**Chen, Q.,** Ding, S., Dai, H., Zhao, T., and Xu, Z.

Poster-Tuesday-pm-s50

Jiangxi Normal University, Tianjin, China (chenqing0701@163.com)

3-Parameter Logistic Graded Response Model

The primary goal in modern IRT research is to expand the existent class of models to cover response data from any "natural" test format. To realize this goal, deeper insights into response processes and the precise way item properties and human abilities interact are needed (van der Linden et al.1997). Although models would not be attached to real test fully, but only by developing and exploring new models continuously, the gap between them can be dwindled and IRT would be perfected. So far, for polytomous item, only item discrimination and grade difficulty have been considered. But in real examination, polytomous items may also have guessing. So in this study, based on Samijima

## ABSTRACTS OF THE CONTRIBUTIONS

---

graded response model, guessing parameter was involved in polytomous response model and a three-parameter graded response model(3PL-GRM)is presented, other works related to this model also be conducted as bellow: 1. Describe 3PL-GRM information function and compare it with GRM information, it is proved that ignoring guessing parameter of polytomous may lead to inaccurate estimation of ability parameters; 2. Based on MMLE/EM algorism, a related program for estimating item parameters in 3PL-GRM is developed.

**Chen, Q.**, Kwok, O.-M., and Luo, W.  
Texas A&M University, College Station, TX (qchen@tamu.edu)

Poster-Tuesday-pm-s50

### What Will Happen If Ignoring a Level of Nesting Structure in Multilevel Growth Mixture Modeling?

Growth Mixture Modeling (GMM, Muthén, 2004) provides a flexible framework for analyzing longitudinal data by combining the use of continuous and categorical latent variables. However, when researchers analyze their data using GMM, they may assume that the participants are independent from each other even though it may not always be true (e.g., D'Angiulli, Siegel, & Maggi, 2004). The topic of Multilevel Growth Mixture Models (MGMM) is relatively new (Asparouhov & Muthén, 2006; Palardy & Vermunt, 2007), and the impact of ignoring a level in MGMM has not yet been well-examined. To extend the findings by Chen, Kwok and Luo (2007) based on cross-sectional models, we examined the impact of ignoring a higher nesting level in MGMM on the accuracy of classification of individuals, and the accuracy as well as the test of significance (i.e., Type I error rate and statistical power) of the parameter estimates for each subpopulation model. Factors including sample size of the subpopulations, magnitude of the average growth trajectory, and magnitude of the between- and within-level variance-covariance matrices for different subpopulation were considered. Implications of the findings are discussed with respect to the application facet.

**Cheng, Y.**  
University of Illinois at Urbana-Champaign, Champaign, IL  
(ycheng6@uiuc.edu)

TS-Wednesday-am-s66

### Classification Accuracy and Consistency under IRT Framework and Optimal Cut-Score Setting

Having a test that enables accurate and consistent classification of test takers is crucial to subsequent decision-making. It is not an easy task, however, to evaluate how accurate and consistent the classifications are made. For example, to determine whether the classifications are accurate, we need to have examinees' true proficiency levels, which in reality are not known. Therefore, researchers developed many methods to "estimate" classification accuracy and consistency rate (e.g., Huynh, 1976a, 1976b; Lee, Hanson & Brennan, 2002; Livingston & Lewis, 1995; Subkoviak, 1976). Many of these papers deal with the cases where tests are scored on number-correct scale. It is a well-known fact that the number-correct score is not a sufficient statistic for the underlying latent trait when 2PL or 3PL model or polytomous IRT models are used. Many testing programs, however, score tests on the basis of these models, and it is therefore interesting and necessary to define and investigate classification accuracy and consistency in the IRT framework. In this paper, we provide a scheme in which classification accuracy and consistency indices can be estimated within the IRT framework, where the cut scores are aligned on the latent trait scale as well as the examinees' trait estimates. We also prove that such classification accuracy and consistency indices will be transformation-invariant under certain regularity conditions. After that, we discuss how to establish the "optimal" cut-score under various criteria.

**Chiu, C.Y.**, and Douglas, J.  
University of Illinois at Urbana-Champaign, Champaign, IL (chiu2@uiuc.edu)

CS-Monday-pm-s21

### Cluster Analysis for Cognitive Diagnosis: A Robustness Study in Relation to Model Misspecification

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

An asymptotic theory has been developed and proved with the goal of using cluster analysis as an alternative to classify data under some given cognitive diagnosis model (Chiu & Douglas, submitted). By summarizing the data with a particular vector of sum-score, K-means cluster analysis or hierarchical agglomerative cluster analysis can be applied with the purpose of clustering subjects who possess the same skills. Following up the theory, the first aim of the current study is to observe the behavior of data in classification when model is misspecified. Then the robustness of the sum-score vector to a variety of cognitive diagnosis models is investigated, which will lead to the discussion on whether it is a benefit by using the cluster analysis to classify data when the true cognitive diagnosis model is unknown. A theoretical investigation for the adequateness of the sum-score to models with different natures is provided. Simulation results are also reported and studied.

**Cho, S.-J.**, and Rabe-Hesketh, S.  
University of California, Berkeley, CA (sjchoo306@gmail.com)

CS-Tuesday-pm-s45

### Alternating Imputation Posterior Estimation of Models with Crossed Random Effects

Latent variable models for categorical responses are difficult to estimate if the latent variables (or random effects/parameters) vary at non-nested levels, such as elementary schools and high schools or persons and items. Clayton and Rasbash (1999) suggested an Alternating Imputation Posterior (AIP) algorithm. For example, in item response models with random item effects, the algorithm iterates between an item wing in which the item mean and variance are estimated for given person effects and a person wing in which the person mean and variance are estimated for given item effects. The person effects used for the item wing are sampled from the posterior distribution estimated in the person wing and vice versa. Clayton and Rasbash (1999) used marginal quasi-likelihood (MQL) and penalized quasi-likelihood (PQL) for estimation within the wings, but this method has been shown to produce biased estimates in many situations, so we use maximum likelihood estimation with adaptive quadrature. For sampling item and person effects, we use either a normal or a discrete approximation for the posterior distributions. We will apply the proposed algorithm to empirical data and present a simulation study with different numbers of items and persons and a range of item and person variances.

**Choi, H.-J.**  
The University of Georgia, Athens, GA (coolchoi@uga.edu)

Poster-Tuesday-pm-s50

### The Impact of Person-Misfit on Scaling and Modeling Students' Growth: Investigating a Bayesian Approach in the Context of a Testlet Model

Monitoring growth in student learning or student progress during school years is a key concern for educators. One aspect of this concern is measuring change in learning over time. In order to infer that change has occurred, a metric needs to be established across time points that enables inferring whether or not change has been positive, negative or neutral. Constructing a common scale for this purpose, in other words, is crucial to making inferences pertaining to students' progress. However, it is a challenging task to construct a valid scale on which to measure and monitor growth across time because ability or achievement is a latent trait. To date, research on growth modeling has frequently been conducted primarily using either structural equation modeling or hierarchical linear modeling. This is in contrast to research on vertical scaling which has been conducted primarily in educational contexts, using either classical test theory or item response theory (IRT). A key component of vertical scaling has been determining the structure of the latent trait and the most appropriate way to equate measures of differing levels of difficulty. The current study intends to implement growth model in vertical scaling via a fully Bayesian approach and investigate impact of person misfit on vertical scaling.

**Choulakian, V.**  
Université de Moncton, Moncton, Canada (vartan.choulakian@umoncton.ca)

CS-Tuesday-am-s37

### On the Rank of Three-Way Arrays

The rank of a three-way array is the number of principal components for three-way data sets in the Candecomp/Parafac. We show that the rank of a three-way array is intimately related to the solution set of a system of polynomial equations. Using this we show some numerical results and we derive bounds on rank values for generic three-way arrays.

**Chow, S.-M.**<sup>(1)</sup>, Allaire, J.C.<sup>(2)</sup>, Hamaker, E.L.<sup>(3)</sup>, and Marsiske, M.<sup>(4)</sup>

IS-Wednesday-am-s52

<sup>(1)</sup> University of North Carolina, Chapel Hill, NC (syymiin@email.unc.edu)

<sup>(2)</sup> North Carolina State University, Raleigh, NC

<sup>(3)</sup> Utrecht University, Utrecht, The Netherlands

<sup>(4)</sup> University of Florida, Gainesville, FL

### Examining Individual Shifts in Learning Dynamics in Group-based State-Space Models

State-space modeling techniques provide a convenient framework for integrating long-term trends and transient intraindividual variability. We generalize an outlier detection procedure proposed by De Jong and Penzer (1998) to a multiple-subject setting and use this approach to diagnose individual-specific shifts in learning dynamics in the context of a group-based state-space model. This model comprises a mixed effects model of learning trends and a stochastic regression model of day-to-day performance variability. By allowing for individualized shocks in the group-based model, we found evidence for an alternative stagewise model of learning in a set of cognitive performance data from  $N = 38$  older adults over 60 days. Other methodological issues and extensions associated with fitting group-based state-space models with individualized shocks will be discussed.

**Clavel, J.G.**<sup>(1)</sup>, and Nishisato, S.<sup>(2)</sup>

CS-Tuesday-am-s35

<sup>(1)</sup> Universidad de Murcia, Spain (jjgarvel@um.es)

<sup>(2)</sup> University of Toronto, Canada

### Total Information Analysis: Applications

This paper presents applications of total information analysis (TIA) to real data and compares our results with those from the traditional analysis based on within-set quantification. In TIA, we will look at the entire space in which both within-set and between-set information spans. Thus, unlike the traditional approach to multidimensional data analysis, TIA deals with all possible dimensions to determine data structure. Our concern is no longer with the principle of parsimony, but with the total information contained in data. Our object is to identify clusters of variables in total space and we will use cluster analysis for this purpose. On the basis of numerical examples, several practical problems and extensions of the procedure to analysis of quantitative data will be discussed.

**Coffman, D.L.**<sup>(1)</sup>, and Blozis, S.A.<sup>(2)</sup>

IS-Monday-am-s5

<sup>(1)</sup> Pennsylvania State University, PA (dlc30@psu.edu)

<sup>(2)</sup> University of California, Davis, CA

### Evaluating Individual Fit in Latent Growth Curve Models

Evaluation of covariance structure models, including latent growth curve models, has almost exclusively focused on overall model fit. Residual diagnostics for evaluating individual fit have not been well developed. The focus here is on detecting outliers by extending measures used in single-level data to longitudinal data. A response may be unusual for a particular individual or an individual may be unusual in comparison to other individuals (i.e., the overall mean). Thus, there are residuals at each level. We simulated data such that a proportion of the individual growth curves was linear and the remaining proportion exponential. We also incorporated measurement error

## ABSTRACTS OF THE CONTRIBUTIONS

---

consistent with a time-varying covariate. We then fit a linear model to all of the growth curves. Hence, the model was misspecified. We computed residuals at both levels and used these to detect individuals for whom the linear model did not fit well. We discuss the standardization of the residuals and their distribution.

**Conijn, J.M.**, van Assen, M.A.L.M., Emons, W.H.M., and Sijtsma, K.  
Tilburg University, Tilburg, The Netherlands (j.conijn@uvt.nl)

IS-Tuesday-am-s33

### Person-Fit Analysis Using Multilevel Logistic Regression

Person-fit research concerns the detection of response patterns that do not fit the hypothesized Item Response Theory (IRT) model or that deviate from the majority of response patterns in a sample. In Reise's (2000) multilevel logistic regression approach to person-fit the probability of a correct response is modeled as a function of the IRT item difficulties. The slope of this function, known as the Person-Response Curve, is assumed to be an indicator of person fit. Deviation from the expected negative slope is indicative of person-misfit. We examine Reise's and alternative multilevel logistic regression approaches for detecting person misfit both conceptually and by means of a simulation study. The performance of the approaches is evaluated under varying conditions like the IRT model underlying the responses, test length, sample size, and type of misfit (systematic misfit, i.e., misfit related to characteristics of individuals, or idiosyncratic person misfit).

**Croudace, T.**

University of Cambridge, Cambridge, United Kingdom (tjc39@cam.ac.uk)

IS-Monday-pm-s16

### Using Psychometric Procedures in Stata to Apply Mokken, Rasch and Logistic IRT Models to General Health Questionnaire Data in the Health and Lifestyle Survey (HALS)

This study examines the traditional [0011] binary scoring of Goldberg's (1972, 1978) General Health Questionnaire (30 item version – GHQ-30) with psychometric procedures available in Stata software. Four models were applied: 1) principal components analysis of the tetrachoric correlation matrix, 2) non-parametric IRT through Mokken's monotone homogeneity model, 3) the Rasch model, and 4) traditional two parameter logistic and probit IRT. Results from the first (n=6000+) and second (n=3000+, 7 years later) of the UK Health and Lifestyle (HALS) survey will be presented. Connections will be made to previous psychometric work using traditional factoring methods. The differences between positively and negatively orientated items will be exploited and new short versions proposed.

**Daniel, R.C.**

Poster-Tuesday-pm-s50

Georgia Institute of Technology, Lawrenceville, GA (rd140@mail.gatech.edu)

### The Effect of Mixed Item Domains on Aptitude Measurement and Response Time

Item modelers have used item response theory (IRT) to make many inferences into the cognitive processes underlying performance on aptitude tests (Embretson, 2001). Recent innovations in item modeling have extended to the point that unproctored home testing is a new trend in testing. These new trends bring new challenges in how to deal with subjects attending to more than one cognitive task. To test the effect of varying cognitive tasks, 3 item domains (Spatial learning, matrix completion, and quantitative reasoning) were chosen and presented both as 3 separate tests and as a single test that change item domain after each question. Performance changed between the two testing formats differently for each item domain.

**Dean, M.J.**<sup>(1)</sup>, and Corter, J.E.<sup>(2)</sup>

IS-Wednesday-am-s60

<sup>(1)</sup>International Baccalaureate, New York, NY (mjd44@columbia.edu)

<sup>(2)</sup>Columbia University, New York, NY

### Cognitive Diagnostic Approaches to Analyzing Performance in the TIMSS (1995) Advanced Mathematics Test

A set of hypothesized content knowledge and process subskills, termed 'attributes', was developed for the Third International Math and Science Study (TIMSS) Advanced Mathematics Test of 1995. A task analysis of test items resulted in a set of twenty-five proposed attributes of three general types: content knowledge, procedural knowledge and specific item type skills. These attributes explained ninety percent of the variance in item proportion correct. These attributes were then used in a Rule Space analysis (and in other cognitive diagnostic testing models) to assess mastery of each attribute for each examinee in the U.S. data and to classify examinees into knowledge states. A high rate of successful classification into knowledge states was obtained. Diagnostic information in the form of attribute mastery probabilities accounted for more than ninety percent of the variance in students' total scores. A residuals analysis was conducted to investigate possibilities for refining the attributes and the underlying cognitive model. Overall, the results suggest that the proposed set of attributes can explain student performance in this test, provide useful diagnostic information, and may lead to a better understanding of mathematical thinking.

**De Boeck, P.**

PA-Wednesday-pm-s70

K.U. Leuven, Belgium (paul.deboeck@psy.kuleuven.be)

### Random Item IRT Models

Typical IRT analyses include person and item modes in the data array (and, of course, there may be more than just two modes). Each of these modes can be modeled with random continuous variables and random categorical variables, or with fixed effects that can also be continuous or categorical (and there could be hybrid versions of these too). In the common IRT models, the persons are modeled with random continuous effects and the items with fixed continuous effects. In addition, persons are sometimes treated also as random categorical, as a hybrid of random categorical and random continuous, or as fixed effects. Applying the random alternatives for fixed effects to items, yields an extension of item response models into unexplored, and not really exploited, kinds of models. The resulting category of *Random Item IRT Models* will be discussed in three ways, in order to illustrate the potential of the models, using: (1) a comparison of four different Rasch models, stemming from the four combinations of random and fixed persons (RP, FP) with random and fixed items (RI, FI) (RP-FI, FP-FI, RP-RI, and FP-RI Rasch), (2) a presentation of explanatory item response models with continuous random item effects functioning as an error component (e.g., Linear Logistic Test Model with error), (3) the formulation of new models for the detection of differential item functioning based on a concept of "aposteriori anchoring," a kind of anchoring based on the data.

**de la Torre, J.**

CS-Monday-pm-s21

Rutgers, The State University of New Jersey, New Brunswick, NJ  
(j.delatorre@rutgers.edu)

### The Generalized DINA Model

In the DINA model, item  $j$  partitions the  $2^K$  latent classes resulting from  $K$  attributes into two latent groups – group  $\eta_j = 1$  consists of individuals who have all the required attributes, and group  $\eta_j = 0$  consists of individuals who lack at least one of the required attributes. Individuals within the same group are assumed to have the same probability of answering the item correctly. However, even if the attributes have been correctly identified and specified, this assumption may not hold for group  $\eta_j = 0$ . That is, because individuals in this group have varying deficiencies with respect to the attribute specification for item  $j$ , their probabilities of success may not be identical. The generalized DINA (G-DINA) model is proposed in this paper to relax this assumption, and allow the latent classes in group  $\eta_j = 0$  to have different probabilities of answering item  $j$  correctly. Several commonly used cognitive diagnosis models can be shown to be special cases of the G-DINA model. In addition, using the appropriate design matrix, the G-DINA model parameters can be transformed into the log-linear metric. Finally, estimates of the G-DINA model parameters can be easily obtained using EM algorithm.

**de Rooij, M.**

IS-Tuesday-pm-s43

Leiden University, Leiden, The Netherlands (rooijm@fsw.leidenuniv.nl)

### Transitional Ideal Point Models for Longitudinal Multinomial Outcomes

For the analysis of longitudinal data three families of models are generally distinguished: the marginal, the transitional, and the subject specific family. In this paper we will propose a transitional model for the analysis of change in a nominal variable. Such an analysis is often hampered by the dimensionality of the problem. We use multidimensional scaling techniques, more specifically the ideal point model, in order to reduce the dimensionality. The model can handle pure transitional data but also allows for explanatory variables. An empirical example with data from the Dutch parliamentary election study 2002/2003 will be discussed in order to discuss all the virtues of the model.

**Ding, S., You, X., Luo, F., Zhao, T., and Xu, Z.**

Poster-Tuesday-pm-s50

Jiangxi Normal University, Tianjin, China (ding06026@163.com)

### Estimating Item Parameters Adaptive in CAT under the 2PLM

It is a very important issue to build a large-scale, high-quality item bank for application of CAT. Usually, the construction of the item bank is a complex work which demands time and energy with high cost. Moreover, the safety of the items cannot be guaranteed when the response data are collected. If raw items can be inserted in the process of administering CAT and the item parameters are estimated at the same time, it will be significant for the construction of CAT item bank. Dr. Han (IMPS,2007,Tokya) proposed an adaptive method to estimate the difficulty parameters of the raw items under 1PLM. It is not sure that the raw items are all fit the 1PLM. If the raw items fit the 2PLM, how to calibrate the parameters? The new scheme to solve the problem is proposed in the paper. Using Han's method to obtain the initial value of the difficulty parameter, then using the relationship of the initial values of discrimination parameter and difficulty parameter (Baker,1992) to obtain the initial value of discrimination parameter. Then the N-R iteration is employed to estimate the parameters in order to obtain more accurate values of estimation. The results of Monte Carlo simulation show that the scheme works well.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Dogan, E.,** Tatsuoka, K., and Kim, Y. Y.

TS-Tuesday-pm-s48

American Institutes for Research, Washington DC (edogan@air.org)

### A Comparison of Rule Space and Item Response Models in Predicting Item and Test Level Examinee Performance

This study examines how well Tatsuoka's Rule Space Model (RSM) explains item and test level examinee performance relative to 2PL and 3PL IRT models. Data came from a 44-item mathematics test included in the 2004 Turkish University Entrance Exam. First, 15 attributes that underlie the test were determined using written protocols from students ( $n=15$ ) and content experts. Each item was coded according to attribute involvement, creating the Q matrix. Each examinee was classified into knowledge states derived from the Q matrix and their attribute mastery probabilities were computed according to the classification results. Next, examinees' probability of correctly answering each item was computed based on RSM and IRT models. RSM assumes that an examinee can successfully answer an item if s/he has mastered all relevant attributes. Therefore the probability that s/he correctly answers an item is equal to the product of his/her attribute mastery probabilities for the attributes involved in that item. Results indicated that the differences between observed and expected examinee performance at both item and test level are relatively smaller (measured by average absolute difference and RMSE statistics) for RSM compared to IRT models. Implications for validation procedures for diagnostic testing models are discussed.

**Dorans, N.,** and Moses, T.P.

IS-Tuesday-pm-s42

Educational Testing Service, Princeton, NJ (ndorans@ets.org)

### Score Equating: Practical Considerations and True Score Models

Holland (1994) made a distinction between tests as contests and tests as measurements. Every contest has rules that should be understood by all contestants. All contestants want to be treated fairly. With measurement, the emphasis is on reliability and validity. For decades, measurement professionals have been trained to focus on the measurement aspects of tests. Fairness has also been of importance, but usually to a lesser degree than measurement. We discuss distinctions between contests and measurements and maintain that score equating is essentially about ensuring fair contests. We focus on the purpose of score equating and its implications for how to produce fair, practically useful linkings. In the process, we consider the utility of true score linking models for ensuring fair contests and producing satisfactory measurements.

**Dumenci, L.,** and Yates, P.D.

CS-Tuesday-pm-s44

Virginia Commonwealth University, Richmond, VA (ldumenci@vcu.edu)

### Performance of Relative Fit Statistics in Distinguishing Mixtures of a One-Factor Model and a Null Model for Binary Items

With a population consisting of two classes, a one-factor model for binary items (i.e., a two-parameter IRT model) and a null model (i.e., a zero-factor model), the performance of AIC, consistent AIC (CAIC), BIC, and sample-size adjusted BIC (saBIC) was examined as a function of the number of items (5, 10, 20), mixing proportions (10%, 25%, 50%, 75%, 90%), item thresholds for the factor and null models (-2, 0 on a logit scale), and  $N$  (500, 1,000, 3,000) in a factorial layout. In addition to the generating model, four competing models were also fit to the simulation data: a latent class analysis with 2 or 3 classes, and a one-factor model with 1 or 2 classes. Of the five competing models, AIC clearly outperformed CAIC, BIC, and saBIC in correctly identifying the true model under all conditions. Results from the relative bias and coverage statistics were somewhat mixed in terms of the accuracy of the recovered parameter estimates.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Dunn, J.**<sup>(1)</sup>, Finkelman, M.<sup>(2)</sup>, Keller, L.A.<sup>(3)</sup>, Keller, R.<sup>(1)</sup>, Kim, W.<sup>(1)</sup>, and Nering, M.<sup>(1)</sup>

CS-Tuesday-am-s27

<sup>(1)</sup> Measured Progress, Dover, NH (dunn.jennifer@measuredprogress.org)

<sup>(2)</sup> Harvard University, Boston, MA

<sup>(3)</sup> University of Massachusetts, Amherst, MA

### An Investigation of the Robustness of the Test Characteristic Curve Mapping Method of Theta Estimation

In the United States, testing for elementary and secondary students has taken on higher stakes as a result of the No Child Left Behind legislation. Many of these assessment programs rely on the test characteristic curve (TCC) mapping method for student ability estimation. This method involves estimating a student ability parameter (theta) by constructing a TCC based on the Item Response Theory (IRT) parameters, and then converting the observed score for each student to a theta estimate through the TCC. This study investigated the impact of various threats to the integrity of a testing program on the quality of the ability estimates. Additionally, the impact of the same threats on Expected A Priori theta estimates were examined to determine if the TCC mapping method provided adequate robustness to the threats when compared with a method that provides more accurate theta estimates. The threats investigated included: differential item functioning (DIF), Local Item Dependence (LID), and multidimensionality. The evaluation criteria were decision accuracy and bias of the theta estimates.

**Edwards, M.C.**

SA-Monday-am-s1

The Ohio State University, Columbus, OH (edwards.134@osu.edu)

### Item Factor Analysis: Where we've Been and Where we Might Be Going

Most of original developments in factor analysis were for scores from experiments or questionnaires (this is true of much of Spearman's and Thurstone's original work). In these cases, a linear relationship between the latent variables and observed scores was often plausible. As time passed, researchers became increasingly interested in conducting factor analyses on item level data. In fact, in many circles, this has become the dominant form of factor analysis. Item factor analysis (IFA), can provide valuable evidence during scale construction regarding dimensionality and item performance, as well as other useful pieces of information. Unlike the sorts of measured variables originally used in factor analytic work, item-level data are often categorical in nature. This violates many of the assumptions that are traditionally made by the model and estimation method. Initial efforts to overcome these difficulties emerged in both the structural equation and item response frameworks. Through much of the 20<sup>th</sup> centuries these solutions were quite different, despite the underlying links between the models. In the past decade, the line between these two frameworks has become increasingly blurry. In this talk I will review the historical development of IFA, provide a snapshot of where we stand today, and discuss where we might find ourselves a decade from now.

**Embretson, S.E.**, and Yang, X.

CS-Monday-pm-s21

Georgia Institute of Technology, Atlanta, GA

(susan.embretson@psych.gatech.edu)

### Multicomponent Latent Trait Models for Cognitive Diagnosis

This paper extends the multidimensional latent trait model (MLTM; Whitely, 1980; Embretson, 1991; Embretson & Yang, 2006) to cognitive diagnosis. MLTM is a non-compensatory item response theory model that predicts the probability of a response from a series of components in the task. This paper extends MLTM to cognitive diagnosis (MLTM-D) by allowing a varying component structure between items. Underlying each component difficulty is one or more features or attributes that may be specified for each item. Simulation data are presented to show parameter recovery. An application to a mathematics tests is presented to illustrate how the model can be used for diagnosis. One advantage to MLTM-D is that it can relate examinee performance to complex attributes and skills with a small number of examinee competencies as compared to the latent class approaches to cognitive diagnosis.

**Emons, W.H.M.**

IS-Tuesday-am-s33

Tilburg University, Tilburg, The Netherlands (w.h.m.emons@uvt.nl)

### Detection and Diagnosis of Person Misfit from Patterns of Summed Polytomous Item Scores under Parametric and Nonparametric IRT Models

This presentation deals with person-fit analysis of polytomously item scores to detect unusual patterns of sum scores on subsets of items. This approach allows a diagnostic approach in which specific hypotheses of person misfit can be tested. In a simulation study, the false-positives rates and detection rates were investigated under varying test and item characteristics, and different types and levels of aberrant response behavior. The performance of the sum-score based approach was compared with the performance of the  $l_{\frac{P}{2}}$  person-fit measure. The simulations showed that the person-fit analysis based on sum-score patterns is useful and performed best for detecting aberrant response behavior that manifests itself locally in the pattern, such as careless responding to reverse-worded items, and traitedness on specific content domains. The person-fit measures discussed are illustrated using real data from personality questionnaires.

**Erosheva, E.A.**, Telesca, D., Matsueda, R.L., and Kreager, D.  
University of Washington, DC (elena@stat.washington.edu)

IS-Monday-am-s6

### Unimodal Hierarchical Curve Registration for Count Data: Analyzing Longitudinal Crime Patterns

A major aim of longitudinal analyses of life course data is to describe the within- and between-individual variability in a behavioral outcome, such as crime. Models for such data typically draw on mixed-effects growth and mixture models. One common method assumes a mixture of groups – defined by distinct polynomial relationships between age and behavior – and incorporates individual-specific polynomial random effects such as age or age squared. We develop an alternative model of life course crime data that assumes a natural age-crime curve, common for all individuals, and allows expected individual crime trajectories to differ by latent patterns of temporal misalignment and scalar amplitude parameters. We analyze self-reported counts of yearly marijuana use from the Denver Youth Survey. Assuming a Poisson distribution for the outcome and a unimodal natural age-crime curve, we estimate the model with Bayesian hierarchical curve registration methods. We take a non-parametric approach to model the natural age-crime curve, and individual time- transformation functions that synchronize individual expected trajectories. Our approach captures individual heterogeneity in meaningful terms by allowing differences in the level of offending and in the timing of features such as age at desistance. We obtain predicted individual crime trajectories that fit the data remarkably well.

**Fargo, J.D.**, Voge, N., and Bohn, C.M.

Poster-Tuesday-pm-s50

Utah State University, Logan, UT (jamison.fargo@usu.edu)

### Advances in the Identification of Sexual Offender Subtypes through Finite Mixture Modeling: Child and Adolescent Victims

Previous research has not empirically identified subtypes of adult male sexual offenders who perpetrate against children and adolescents. Development of a reliable and ecologically valid taxonomy has potential to guide prevention/intervention efforts. To this end, a series of (1-10 class) finite mixture models was tested using data from a sample of 455 adult male sexual offenders whose victims were children and adolescents. Data were obtained from a nationally representative, 2-stage cluster sample of US state and federal prisoners. Class indicators included measures of socio-demographic characteristics, criminal histories, developmental experiences, substance use/abuse, and victim and sexual offense characteristics. Results indicated a 4-class model possessed the best fit to the data when compared to competing models following bootstrapped -2LL difference tests. The four classes were identified as 1) high-SES, first-time offenders (30%); 2) moderate-SES, criminal recidivists with a high degree of substance abuse and personal sexual/physical abuse, having families of their own (mostly divorced), victimizing both (their

## ABSTRACTS OF THE CONTRIBUTIONS

---

own) children and adolescents well known to them (29%); 3) moderate-SES, all criminal recidivists, who mostly victimized children, with low personal sexual/physical and substance abuse history (21%); and 4) low-SES, anti-social, violent criminals who victimized mostly a mix of stranger and well known adolescents (20%).

**Feldman, B.J.**, and Masyn, K.E.

IS-Tuesday-am-s34

<sup>(1)</sup> University of California, Berkeley, CA (b.feldman@berkeley.edu)

### Measuring and Modeling Adolescent Alcohol Use

Adolescent alcohol use is not straightforward to measure or model longitudinally. There is considerable difference between the time scale of the behavior (weekly, daily, or even hourly) and that of assessment (usually yearly), so self-report data are generally imprecise and may not be very accurate. The data are typically categorical and often highly skewed, requiring special modeling techniques. Finally, for appropriate models, researchers generally lack good measures of fit, making model selection somewhat tenuous. In this study, we simulate five years (1825 days) of daily drink counts, using a population model based on typical adolescent drinking trajectories and the general day-to-day intermittency of drinking. With the resulting data, we calculate annual summary measures: categorical summaries of drinking frequency; counts of typical drinks-per-drinking-occasion; and frequency-quantity measures, calculated by multiplying frequency by typical amount drunk. Both continuous and categorical longitudinal models are fit to the resulting data (e.g., random-effects models, growth mixture models, two-part models), and the results of the analyses are compared with the simulated data, on individual and population levels, with respect to variability and patterns of change. We investigate how measurement and analytic approaches may act and interact to affect the accuracy of models of change in alcohol use.

**Finch, H.**<sup>(1)</sup>, and French, B.<sup>(2)</sup>

CS-Wednesday-am-s54

<sup>(1)</sup> Ball State University, Muncie, IN (whfinch@bsu.edu)

<sup>(2)</sup> Purdue University

### Comparison of MANOVA and a Structural Equation Modeling Method for Comparing Observed Score Group Means on Multiple Dependent Variables: A Simulation Study

Multivariate analysis of variance (MANOVA) is a popular approach for simultaneously comparing multiple means across groups. Several studies have demonstrated limitations of this approach when assumptions are violated, particularly that of homogeneous covariance matrices. An alternative approach to these multivariate comparisons of means described by Raykov (2001) has been proposed based upon structural equation modeling (SEM). SEM does not require the assumption of equal covariance matrices, and thus may be preferred to MANOVA in such cases, particularly with large samples. Little empirical evidence under controlled conditions (i.e., Monte Carlo simulation studies) exists to support if the benefits of SEM in this context is realized in practice. The current study compared the Type I error and power rates of the MANOVA and the SEM approach for observed variable means, across varying (a) sample sizes, (b) covariance equality, (c) group size ratios, (d) distributions, and (e) numbers of variables. Simulation results suggest that with total sample sizes equal to or greater than 100, the SEM approach maintained the nominal Type I error rate and had higher power compared to the MANOVA approach. Furthermore, when covariance matrices were unequal, the SEM approach outperformed (i.e., lower Type I error, greater power) the MANOVA approach.

**Finkelman, M.**<sup>(1)</sup>, Kim, W.<sup>(2)</sup>, and Roussos, L.<sup>(2)</sup>

TS-Tuesday-pm-s48

<sup>(1)</sup> Harvard University, Boston, MA (mfink@jimmy.harvard.edu)

<sup>(2)</sup> Measured Progress, Dover, NH

### Test Assembly for Cognitive Diagnosis Models using Attribute-Level Information

## ABSTRACTS OF THE CONTRIBUTIONS

---

Much recent psychometric literature has focused on cognitive diagnosis models (CDMs), a promising class of instruments used to measure the strengths and weaknesses of examinees. In order to achieve the maximal discriminatory power possible from a CDM, it is crucial to employ automated test assembly (ATA) methods that make the most effective use of item pools. This article introduces two new ATA methods, along with several complementary variations, that utilize the attribute-level CDM discrimination indexes of Henson, et al. (in press). Our new procedures, called the maximin and weighted maximin methods, are designed to provide satisfactory overall information as well as adequate discriminatory power along each attribute. Specifically, they aim to deliver low, approximately uniform error rates across the  $K$  attributes of interest. The newly introduced procedures are thus conceptually different from an existing ATA index, the CDI, which measures overall rather than attribute-level item information. Methods are compared under multiple simulation conditions.

**Fox, J.-P.**

CS-Tuesday-am-s30

University of Twente, Enschede, The Netherlands (j.p.fox@utwente.nl)

### Advanced Posterior Predictive Assessment

A standard Bayesian statistical tool for model checking is based on a discrepancy measure to investigate the compatibility of the model with the data. The reference distribution of the discrepancy measure is used to measure the extremeness of the observed discrepancy. The predictive diagnostic checks utilized via  $p$ -values are specifically important when no fully specified alternative model is available. A common problem with the computation of posterior predictive  $p$ -values (ppp-values) is its double use of the observed data, first to compute the posterior distribution of the model parameters for determining the posterior predictive distribution, and second to compute the ppp-value. This double use of the data leads to unnatural behavior of the ppp-value. The standard approach for obtaining ppp-values is extended in two different ways to obtain Uniform distributed unbiased  $p$ -values. In the first approach, it is shown that the marginal distribution of the data can be used to calibrate the ppp-value. In the second approach, the reference distribution of a conditional posterior predictive check is determined that leads to a conditional ppp-value. The corresponding marginal ppp-value can be obtained via MCMC. Several examples are given in the context of IRT and nonlinear mixed models.

**France, S.L.,** and Carroll, J.D.

CS-Monday-am-s7

Rutgers University, Newark, NJ (sfrance@andromeda.rutgers.edu)

### Distanced Based Metric Multidimensional Scaling for Complex Data

Distance based metric multidimensional scaling (DMMDS) was first proposed by Kruskal in 1964. Initially, DMMDS was typically applied to small data sets from psychological experiments. Modern psychological research deals with large and complex data, such as test data, MRI data, and demographic data. Our work concentrates on the development of DMMDS as a serious tool for the visualization and interpretation of large datasets. The problems encountered in developing MDS algorithms for the data found in large scale, realistic data are numerous. Underlying object $\times$ stimuli data may be of different measurement types. As the number of items to be visualized grows, then the computational and memory requirements of DMMDS also grow. We discuss methods for calculating one or more distance/dissimilarity or other proximity matrices for these data, and for combining multiple such matrices into a single distance-like matrix to which DMMDS is applied. We pay particular attention to scaling using the  $L_1$  metric. A rationale for the use of the  $L_1$  metric rather than the  $L_2$  metric when scaling data of high dimensionality is given by recent research in data mining that shows that as the dimensionality of data increases, so the concept of distance becomes increasingly ill defined in Euclidean space.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Fukunaka, K.**, and Toyoda, H.  
Waseda University, Tokyo, Japan (tarotwork@ruri.waseda.jp)

Poster-Tuesday-pm-s50

### Generation of Direct Graphs and Chain Graphs by Using SEM

In this presentation, it is shown that direct graphs and chain graphs of graphical modeling (GM) can be expressed by the framework for SEM (Study 1), and methodology to be applied to factors is proposed (Study 2). In Study 1, I will describe method and procedure to construct direct graph or chain graph of the framework for SEM. I reanalyzed some data which already analyzed by the GM using this methodology, which was in excellent agreement with the values by the previous study. On the other hand, the method in study 1 is extended the method using factors in study 2, and it is shown that this methodology can be applied to the fields of educational psychology. Concretely speaking, focusing on the situation that college students who is majoring in psychology learn statistics, I made exam questions about statistics and questionnaire factors are "generalized self-efficacy", "self-regulated learning", "test anxiety", "task-specific self-efficacy" and "test performance" and carried out it. This result was constructed of model by the chain graph. In the result, I could make a model for self-efficacy and learning of statistics to college students who is majoring in psychology.

**Garcia, R.**, and Kenny, D.A.  
University of Connecticut, Storrs, CT (randigarcia@gmail.com)

Poster-Tuesday-pm-s50

### Power Differences in Testing Fixed and Random Effects

It can happen in a study that a given fixed variable, e.g., gender of perceiver, has a statistically significant effect, but the random effect (e.g., perceiver) does not have a statistically significant effect. A parallel finding occurs in some meta-analyses, when studies do not vary more than expected by sampling error, yet particular types of studies differ from each other (e.g., more recent studies have larger effect sizes than older studies). The general phenomenon is that tests of a fixed effect appear to have more statistical power than tests of a random effect. We are left with the anomaly that there is an effect on something that does not vary. We conducted a simulation in which we created a dichotomous fixed effect which is nested within a random effect and we calibrated the effects to be of equal size. We show with this simulation, that the power of the test of the fixed effects is much greater than the test of random effects. Moreover, we show that the two tests are negatively correlated increasing the probability that the fixed effect is statistically significant but the random effect is not statistically significant.

**Geerlings, H.**, van der Linden, W.J., Glas, C.A.W., and Holling, H.  
University of Twente, Enschede, The Netherlands (h.geerlings@utwente.nl)

CS-Tuesday-pm-s45

### Modeling Rule-Based Item Generation

One of the advantages of rule-based item generation is a relatively large number of items generated through the application of only a few rules. Families of items are created through combinations of these rules, each of which influences the difficulty of all items within a family. Surface features of the items within a family are changed to ensure that the items do not look the same. A hierarchical model will be presented that can account for the fact that items are grouped in families and that families are created through the application of rules. The first-level model is the three-parameter normal ogive model, which describes the relationship between the ability and item parameters in predicting the probability of a correct response. At the second-level, the (transformed) item parameters are modelled as a multivariate normal distribution with family means and covariances. The family mean of the difficulty parameter is represented by a linear combination of the effects of the rules applied for that family. Every item then has the same statistical properties of the family it belongs to, plus a random component. The parameters of the model are estimated in a Bayesian framework, using a data-augmented Gibbs sampler.

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

**Gonzalez, E.J.**

IS-Wednesday-am-s60

IEA Data Processing Center, Hamburg, Germany (egonzalez@ets.org)

### Current Operational Analyses of Large Scale Survey Data and Publicly Available Databases

The models and approaches used to analyze data from large scale educational surveys are based on extensions of item response theory to models with a more complex population structure. More specifically, the models use a latent regression to describe the distribution of the latent proficiency variables from the IRT model in terms of conditional distributions based on the available background data. Data from the context questionnaire are aggregated and preprocessed before they enter the latent regression, and the data products derived from the models used are commonly referred to as plausible values. Technically, these plausible values (PVs) are imputations from the posterior distribution of proficiency given item responses and background data. Data products that contain records for each participant contain PVs as well as weights for each examinee and the responses to the context questionnaire, and possibly variables from teacher or school questionnaire repeated in the files. The talk will give an overview of current operational procedures and the derived data products, as well as resources for requesting data and online tools to tabularize data from large scale surveys on the World Wide Web.

**Goodrich, B.**

CS-Wednesday-am-s55

Harvard University, Cambridge, MA (goodrich@fas.harvard.edu)

### Semi-Exploratory Factor Analysis and Associated Software

This January, I released open-source software to estimate EFA and CFA models via a genetic algorithm, which includes a new estimator called semi-exploratory factor analysis (SEFA) that reflects a new way of thinking about factor analysis. My goals for IMPS are to introduce SEFA to the factor analysis community and to invite anyone who is interested in any aspect of factor analysis or SEM to contribute code to this project. SEFA requires the analyst to specify how many coefficients for each factor must be exactly zero but — unlike CFA — does not require the researcher to specify which coefficients are zero. To overcome rotational indeterminacy the required number of zeros per factor must be at least one less the number of factors. SEFA uses a genetic algorithm to minimize the discrepancy function over the locations of these exact zero coefficients and the values of the corresponding non-zero parameters, potentially subject to a wide range of non-linear inequality restrictions that the analyst may impose on functions of parameters. Many of the available restrictions are suggested by Allen Yates in his 1987 book on EFA, which is an under-appreciated work whose lessons can be extended to SEFA and CFA models.

**Grelle, D.**

Poster-Tuesday-pm-s50

PreVisor and University of Georgia (dgrelle@previsor.com)

### Criterion Dynamism and Growth Mixture Modeling: Exploring Selection Assessment Utility by Identifying Latent Classes of Performance Change over Time

There is growing consensus among organizational researchers that job performance changes over time, especially in jobs requiring skill acquisition. The nearly ubiquitous finding of declining correlations between selection assessments and performance as time increases calls into question the utility of selection tools. Studying performance cross-sectionally or at the mean level cannot address why predictive validity decreases over time. Latent growth models assume that growth trajectories come from a single population with normal variance around the parameters, but this is likely a naïve assumption. Growth mixture modeling (GMM) does not make this assumption and was used to identify classes of performance change that would be considered desirable by organizations' stakeholders. Objective job performance data for three metrics were collected over nine months for a sample of 203 call center agents. Class membership probabilities were calculated using Mplus (Muthen & Muthen, 2008) for a model in which individual performance trajectories were estimated with assessments of cognitive ability, emotional resilience, sales ability, and conscientiousness as covariates. Across the three metrics, GMM identified three to four classes of

## ABSTRACTS OF THE CONTRIBUTIONS

---

growth. The assessments successfully predicted membership in "desirable" trajectories, suggesting that the more holistic approach of GMM is appropriate for assessing the utility of selection assessments.

**Griffiths, T.**

KN-Tuesday-am-s24

University of California, Berkeley, CA (tom\_griffiths@berkeley.edu)

### The Modern Intuitive Statistician

The idea that human inferences can be characterized by the principles of Bayesian statistics has been championed and derided at various points in the history of cognitive science. In this talk, I will argue that this idea is still valuable in making sense of one of the most impressive aspects of human reasoning: our ability to solve inductive problems, such as learning causal relationships or learning categories based on only a few observations. In these cases, Bayesian statistics gives us a tool for understanding why people reach the conclusions that they do, and for identifying the assumptions that guide their inferences. In particular, I will focus on how tools from modern Bayesian statistics, such as hierarchical Bayesian inference, nonparametric latent variable models, and Markov chain Monte Carlo, can be used to shed light on some of the sophisticated inferences that people make in everyday life.

**Gueorguieva, R.**

IL-Tuesday-pm-s41

Yale University School of Medicine, New Haven, CT  
(ralitza.gueorguieva@yale.edu)

### Modeling Longitudinal Trajectories Using Growth and Growth Mixture Models

Recent advances in longitudinal statistical modeling allow realistic description of patterns of change over time, use of all available data on an individual and assessment of the effects of both stationary and time-dependent variables. Traditional growth modeling assumes that every individual follows the same type of trajectory over time, while growth mixture models allow data-driven identification of distinct classes of developmental trajectories within a population. The latter models also allow characterization of the individuals most likely to belong to each class, assessment of treatment or intervention effects on trajectory membership and simultaneous modeling of trajectories of related behaviors. In this presentation we will describe trajectory-based approaches, will discuss their advantages and disadvantages, and will formulate recommendations for the implementation of such models. We will use examples from randomized clinical trials in alcohol dependence to illustrate identification of patterns of change over time, selection of number of trajectory classes, assessment of treatment effects and effects of time-dependent covariates, and joint modeling of treatment and compliance trajectories. Software for fitting such models will also be briefly discussed. [Supported by the Department of Veterans Affairs Cooperative Study Program, the Center for Translational Neuroscience of Alcoholism (P50 AA012870- 05).]

**Gundula, A.M.**

CS-Monday-am-s7

University of Alberta, Edmonton, Canada (gundula@ualberta.ca)

### Investigating the Inter-Rater Reliability of Test Scores: An MDS Analysis

If scores from educational assessment are to be useful and defensible, the inter-rater reliability of the tests scores must be evaluated. A lot of research has been done on evaluating inter rater reliability of test scores but no research has been on evaluating the inter-rater reliability of test scores using Cluster Analysis and Multidimensional Scaling. Five raters were used and their scorings to each and every item were analyzed using multidimensional scaling (MDS) procedure followed by a hierarchical cluster analysis of the MDS stimulus coordinates. In this study, scoring differences were most pronounced when looking at the percentage of question scored exactly the same at specific score points, and the percentage of exact agreement corrected for chance. The findings suggest that MDS analysis can provide substantive information pertaining to the inter-rater reliability of test scores. The advantages and disadvantages of using MDS and cluster analysis with item similarity are discussed.

**Hafdahl, A.R.**

CS-Tuesday-am-s37

Washington University, St. Louis, MO  
(hafdahla@gmail.com or arhafdah@wustl.edu)

### Meta-Analysis for Functions of Heterogeneous Correlation Matrices

Correlation analyses in primary research often focus on certain substantively interesting functions of the correlations among several variables, such as partial, (squared) multiple, or canonical correlations; coefficients in regression, path, factor, or covariance-structure models; eigenvalues; or various combinations of these (e.g., contrasts, ratios). Research synthesists rarely meta-analyze such functions, however, perhaps due to lack of principled methods. In this paper I extend previous work on simpler cases (e.g., 1 study, 1 focal correlation, fixed-effects models) to meta-analytic estimation and inference for popular functions of correlations -- possibly vector-valued -- with particular attention to the case of between-studies heterogeneity. Standard techniques with recent refinements are used to estimate the correlation-matrix parameters' expectation and covariance matrix in the Pearson-r or Fisher-z metric, and an integral transformation of the implied parameter distribution yields the function's mean and covariance matrix. Proposed methods for inference about the function's mean include a delta-method covariance matrix as well as parametric and nonparametric bootstrapping. These approaches are contrasted with two related strategies that entail applying the focal function to either the observed correlation matrices or the mean correlation matrix. Monte Carlo studies of the focal techniques are presented.

**Halpin, P.F.**, and Maraun, M.D.

CS-Tuesday-pm-s44

Simon Fraser University, Burnaby, BC, Canada (phalpin@sfu.ca)

### Empirical Criteria for Model Selection: Linear Factor and Finite Mixture Structures

It is well known that the  $k$ -dimensional linear factor and  $(k + 1)$ -class finite mixture models imply indistinguishable covariance factorizations. The significance of this result is conceptualized in terms of two related topics: model selection and latent parameter estimation. It is argued that, although the latter is well addressed by fitting procedures, the former is best accomplished without direct reference to latent quantities. The identical covariance factorizations thus motivate discussion of novel empirical criteria, based on the functional behavior of observed conditional covariances, for distinguishing between the two types of models. The case of  $k = 1$  is addressed in detail as this solution is entirely tractable and lends itself to various strategies for generalization to higher dimensions. Generalization for the factor model is straightforward, but for the finite mixture model the conditional covariance functions become increasingly unwieldy with the number latent classes. The results given here are nonetheless sufficient to provide sample-based selection criteria that do not require antecedent model specification. These are formulated and compared to extant "non-parametric" model selection procedures.

**Hamaker, E.L.**<sup>(1)</sup>, and Grasman, R.P.P.P.<sup>(2)</sup>

IS-Wednesday-am-s52

<sup>(1)</sup> Utrecht University, Utrecht, The Netherlands (e.l.hamaker@fss.uu.nl)

<sup>(2)</sup> University of Amsterdam, Amsterdam, The Netherlands

### How to Study Regime-Switches in Affect

Hidden Markov models are valuable for the study of processes that are characterized by regime-switches and when it is not clear in which regime the system is at a given occasion. Kim (1994) introduced a state-space model with regime-switching for which he combined the (regular) Kalman filter with the Hamilton filter for the probabilities of switching from one regime to another. In this presentation we extend the Kalman filter and Hamilton filter so it becomes possible to handle missing data. In addition we focus on how to select the appropriate number of regimes. We illustrate this with daily measurements of positive and negative affect in people diagnosed with bipolar disorder and normal controls.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Han, K.T.**, and Wells, C.S.

TS-Monday-pm-s22

Graduate Management Admission Council (khan.educ@gmail.com)

### Impact of Multidirectional Item Parameter Drift on Test Equating and Proficiency Estimates

Despite an extensive number of studies that have examined methods for detecting items exhibiting item parameter drift (IPD), the question that has not been addressed yet is how much IPD is needed until the effect is consequential. Han and Wells (2007) studied the impact of IPD on test equating and proficiency estimates in the context of unidirectional IPD (i.e., all IPD items were simulated to be easier at Time 2). However, in a large-scale assessment it is not unusual to observe multidirectional IPD in which some items become harder while other items become easier. Thus, it is important to examine how multidirectional IPD influences test equating and proficiency estimates. Therefore, in this study, the impact of different combinations of linking items with various multidirectional IPD on the test equating procedure will be investigated for three popular scaling methods (mean-mean, mean-sigma, and TCC method) via a series of Monte Carlo simulation studies. It is hypothesized that multidirectional IPD will influence the amount of random error observed in the linking while the effect of systematic error will be minimal.

Hancock, G.R.<sup>(1)</sup>, **Choi, J.**<sup>(2)</sup>, and Kroopnick, M.H.<sup>(1)</sup>

CS-Monday-pm-s19

<sup>(1)</sup> University of Maryland, College Park, MD (ghancock@umd.edu)

<sup>(2)</sup> The George Washington University

### Simplified Sample Size Determination for Two-Point Repeated Measure Structured Means Modeling on a Single Latent Construct

Structured means modeling for repeated measure (within-subjects) designs is useful for researchers interested in assessing mean-level differences for a population of individuals (or for populations of matched individuals) on a latent construct under multiple conditions or over time. By far the most common example is a pre-post design, in which change in the mean level of a latent construct before and after treatment is being evaluated; the benefit of the latent aspect of this approach is the control of measurement error within the evaluation process. For this common two-point ( $J=2$ ) repeated-measure case, the current presentation will derive a simplified process for sample size determination, relating necessary sample size to a measure of latent effect size and to an increasingly common measure of construct replicability, maximal reliability. The issue of correlated error structures is addressed for the homogeneous case, and challenges associated with more complex multi-point designs (i.e.,  $J>2$ ) are also discussed.

**Harring, J.R.**

IS-Monday-am-s5

University of Maryland, College Park, MD (harring@umd.edu)

### Modeling Nonlinear Change in Latent Variables over Time

The nonlinear mixed effects (NLME) model for continuous repeated measures data has become an increasingly popular and versatile tool for investigating nonlinear longitudinal change in observed variables. As an extension of this framework, this research considers a NLME model for describing nonlinear change of a latent construct over time, where the latent construct of interest is measured by multiple indicators gathered at each measurement occasion. To accomplish this, the nonlinear mixed effects model is modified to include a measurement model that explicitly expresses the relation of the observed variables to the latent constructs. An efficient method for maximum likelihood estimation of the model is developed so that the dimensionality of integration is reduced to the number of nonlinear coefficients. An example using education data will be provided to illustrate the utility of the model.

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

**Harris, D.**

IS-Tuesday-pm-s42

ACT, Inc., Iowa City, IA (deborah.harris@ract.org)

### Linking Across Forms in Vertical Scaling

Vertical scaling refers to the process of linking different levels of an assessment, which measure the same construct, onto a common score scale; elementary and secondary test batteries report scores on a vertical scale, and much research has been on various aspects of creating a vertical scale, including definition of growth, scaling design, statistical methods, type of scales, etc. (see Harris, 2007; Harris, Hendrickson, Tong, Shin, & Shyu, 2004; Kolen, 2003). One issue that has not been addressed much in the literature is that of maintaining vertical scales over time, and over new forms (Hoskens, Lewis, and Patz, 2003 is an exception). For example, should new “grade 3” forms be equated to the original “grade 3” form, or should there be an attempt to link the entire range of, say, grade K to grade 8 forms to the original set of forms the scale was set on? What types of drift, or error, are we apt to see over time? How often should a vertical scale be “realigned”? This study looks at theoretical issues involved in maintaining a vertical scale across forms, such as what are appropriate criteria to use to evaluate the effectiveness of an equating, using real and simulated data.

**Hayashi, K., Shimizu, Y., and Kano, Y.**

CS-Wednesday-am-s65

Osaka University, Osaka, Japan (khayashi@sigmath.es.osaka-u.ac.jp)

### Penalized Boosting Algorithm for Mislabeled Data

Mislabeled data can distort any statistical classification results. In this talk, a new boosting method is proposed in the case where mislabeled subjects are included in a training data set. The true label is treated as a latent variable, and the risk function is defined as the mean of the loss for true labels conditioned on observed data. The new method is a penalized boosting algorithm based on a model for the relation between unknown true labels and observed data. Penalized boosting algorithm contributes to avoiding the overlearning since there is the trade-off between the goodness of fit of the classifier and the stability thereof. It is shown that the penalized boosting algorithm developed here also circumvents the distortion caused by mislabeled subjects. Numerical experiments are conducted to evaluate prediction performance of the proposed method for true labels.

**Heo, M.**

Poster-Tuesday-pm-s50

University of Georgia, Athens, GA (msnica@uga.edu)

### Science Motivation and Epistemological Beliefs of South Korean High School Students Who Aspire Toward Science

This study explores relationships among the science motivation and epistemological beliefs of academically advanced Korean high school students who aspire toward the sciences. One hundred and seventy-two tenth grade students who attended science high schools in Seoul, Korea, completed a questionnaire adapted from the Motivational Strategies for Learning Questionnaire (MSLQ: Pintrich, Smith, Garcia, & McKeachie, 1993) and Epistemic Belief Inventory (EBI: Schraw, Bendixen, & Dunkle, 2002). The schools are for students who excel in science and math. To explore the factor structure of each scale, exploratory factor analyses were conducted. The analyses supported the five factor structure of motivation (Efficacy/Challenge Seeking, Performance Goals, Value, Control of Learning Beliefs, and Test-Anxiety) and the five factor structure of epistemological beliefs (Simple Knowledge, Certain Knowledge, Fixed Ability, Quick Learning, and Omniscient Authority). Using Lisrel 8.54, a path model was established that explains the relationships among the factors.

**Hessen, D.J.**

CS-Tuesday-am-s29

Utrecht University, Utrecht, The Netherlands (D.J.Hessen@uu.nl)

### Parameter Estimation and Likelihood Ratio Tests for Parametric Constant Latent Odds-Ratios Models

In this presentation, the focus is on parameter estimation and goodness-of-fit assessment under parametric constant latent odds-ratios models (CLORs) for dichotomously score items. CLORs models have been proposed as alternatives to traditional models, such as the Rasch model and the two- and three-parameter logistic Birnbaum models. Special attention is given to three special CLORs models. For the estimation of the structural parameters of the models a marginal maximum likelihood (MML) procedure is proposed. A Bayesian approach is considered for person parameter estimation. Likelihood ratio tests for assessing the overall and relative goodness-of-fit of specific CLORs models are presented. For assessing the overall goodness-of-fit of a specific CLORs model, the usual likelihood ratio test against the saturated multinomial model can be used. For assessing the relative goodness-of-fit of two CLORs models a chi-square difference test is proposed. Finally, all procedures considered are demonstrated in a real data example.

**Ho, A., Furgol, K., and Magda, T.**

TS-Wednesday-am-s66

University of Iowa, Iowa City, IA (andrew-ho@uiowa.edu)

### Using Parametric Assumptions to Frame Trends in Categorical Proficiency Data Under the U.S. No Child Left Behind Act

Many states and countries summarize large-scale test scores as percentages within or above score categories such as Basic, Proficient, and Advanced. Changes in these percentages are widely interpreted as trends. These percentage-based trends can be misleading. Trends are expected to be larger for central categories where greater numbers of students are adjacent to category boundaries. When states or countries limit reporting to these categorical data, secondary analysts are faced with impoverished trend estimates that may be deeply misleading as expressions of educational progress. In this paper, we introduce a framework for supplementing this essentially nonparametric data with parametric assumptions. We demonstrate that inverse-normal and log-log transformations are the simplest of a family of procedures that enforce parametric assumptions on nonparametric data, and we introduce other members of this family. Members of these families allow estimation of effect sizes from changes in categorical data under assumptions that distributions are normal, skewed, triangular, uniform, or otherwise. We conclude that this procedure is useful for the cross-state aggregation of trend data when only categorical data are available, and we show that parametric assumptions can set baseline expectations for trends that can afford investigations of the inflation of test scores at particular category boundaries.

**Hofmans, J., Schepers, J., Kuppens, P., and Van Mechelen, I.**

CS-Tuesday-am-s35

K.U. Leuven, Belgium (joeri.hofmans@psy.kuleuven.be)

### Capturing the Nature of Two-Way Interactions: A Two-Mode Clustering Based Approach

When studying the relationship between two categorical predictor variables and a continuous criterion often not only the absence or the presence of an interaction but also the nature of the interaction is of key importance. In this regard, one can distinguish between four types of interactions: interactions that are ordinal or disordinal with respect to none, one or both predictor variables. Recently, Schepers and Van Mechelen (submitted) proposed a two-mode clustering based approach in which these four different types of two-way interactions are imposed on the data by fitting constrained Real-Valued HICLAS models. Subsequently the best model, in terms of predictive power, is selected by comparing the constrained models using a resampling based cross-validation procedure. Such an approach allows to reveal which type of interaction most adequately represents the major pattern of the relationship between both predictor variables and the criterion. In this talk, using three examples from the domain of contextualized emotion psychology, the validity of this approach will be demonstrated by showing that different types of interactions are

## ABSTRACTS OF THE CONTRIBUTIONS

---

found and that each type warrants a different psychological interpretation that is both logically sound and in agreement with the existing emotion literature.

**Holland, P.W.**, and Strawderman, W.  
Educational Testing Service, Princeton, NJ (holland@ets.org)

IS-Tuesday-pm-s42

### How to Average Equating Functions if you Must

The interest in forming averages of equating functions arises in several ways—from combining independent estimates of the same equating function to forming compromises between equating functions computed under different assumptions about anchor tests. The simple point-wise average of two equating function (with weights that are independent of the two functions) will, in general, fail to satisfy the symmetry property that is usually required of equating functions. That is, the average of the inverses of the two equating functions will not necessarily be the inverse of the average of the two functions. We discuss a large family of simple point-wise averages of equating functions (with weights that do depend on the slopes of the two functions) that do satisfy the symmetry property for linear equating functions as well as other natural properties that an average of equating functions ought to possess. Moreover, we propose a general procedure, the symmetric  $w$ -average or *swave*, for computing weighted averages of linear or non-linear equating functions that always satisfies the symmetry property. An example based on real data is given.

**Hong, Y.**, and de la Torre, J.  
Rutgers University, the State University of New Jersey  
(yuanhong@eden.rutgers.edu)

CS-Wednesday-am-s57

### Parameter Estimation with Small Sample Size: A Higher-Order IRT Approach

Sample size ranks as one of the most important factors that affect the item calibration task. However, due to practical concerns (e.g., item exposure) items are typically calibrated with much smaller samples than what is recommended. To address the need for a more flexible framework that can be used in small sample item calibration, this paper proposes an approach pertains to the treatment of the dimensionality of the assessments in the calibration process. This approach is based on the higher-order IRT (HO-IRT) model. The HO-IRT model is a multi-unidimensional model that uses in-test collateral information and represents it in the correlational structure of the domains through a higher-order latent trait formulation. Using MCMC method in a hierarchical Bayesian framework, the item parameters, the overall and domain-specific abilities, and their correlations are estimated simultaneously. The feasibility and effectiveness of the proposed model are investigated under varied conditions in a simulation study, and illustrated using actual assessment data.

**Hong, Y.**<sup>(1)</sup>, and Yao, L.<sup>(2)</sup>  
<sup>(1)</sup> Rutgers University, New Brunswick, NJ (yuanhong@eden.rutgers.edu)  
<sup>(2)</sup> CTB/McGraw-Hill

Poster-Tuesday-pm-s50

### Comparison Among Major Value-Added Models: A General Model Approach

Value-added models (VAM) are becoming increasingly popular within accountability-based educational policies, as they purport to separate out the effects of teacher and schools from background variables. Given the fact that many states have implemented or are planning to implement VAM into their accountability systems and that the consequence of an error impacts students, teachers, and schools, it is crucial to make sure that each model is studied thoroughly and that all of the caveats are clearly identified. In addition, research to date has shown that applying different VAM models produces very different results. We first introduce a general VAM framework, from which all the models can be derived. Statistical and computing specifications are given for each of these models. The models were fitted to obtain value-added measures of school performance by time and subject area, using empirical data set

## ABSTRACTS OF THE CONTRIBUTIONS

---

with three years of math and reading scores. We investigate the differences among these models by comparing their value-added measures, such as the estimated demographic effects, school effect and the school persistence effect. MCMC algorithm is used to implement the Bayesian methods for some of the most complex models.

**Horton, D.**, and Markus, K.

Poster-Tuesday-pm-s50

John Jay College of Criminal Justice, Brooklyn, NY (dhorton79@gmail.com)

### An Examination and Interpretation of the Factor Structure and Validity of the Suicide Prevention Screening Guidelines

Although suicide is a leading cause of death in U.S. jails, the validity of the widely used Suicide Prevention Screening Guidelines (SPSG) has not been adequately researched. The present study applied exploratory factor analysis in a sample of 10,678 inmates from a large urban jail. The SPSG consists of 18 dichotomous items. Comprehensive Exploratory Factor Analysis software (Browne, Cudek, Tateneni and Mels, 2004) was used to factor tetrachoric correlations. Separate analyses were required for men and women. Two items with endorsements linked to other items and one with an extremely low base rate for women had to be removed. Only four of the remaining items exceeded 10% yes responses. Such methodological considerations reflecting the SPSG's dichotomous items and their low base rates were not accounted for in its development. The factor structures produced in the present study differed significantly from the one reported by the test's developers. The most interpretable factor solutions appeared to over-factor, indicating the SPSG should include more items in order to better cover the construct domain. These results highlight the tension between the need for interpretable factor solutions that satisfactorily represent the complexity of suicide risk and the pragmatic need to keep the screening instrument short.

**Hsu, H.-Y.**, Kwok, O.-M., Zou, Y., and Wu, Y.-Y.

TS-Tuesday-am-s31

Texas A&M University, College Station, TX (hsuhy0914@neo.tamu.edu)

### Testing the Effectiveness of Fit Indices in Detecting Misspecification in Multilevel Structural Equation Models: A Monte Carlo Study

In social sciences, data are likely to have a hierarchical (or multilevel) structure. There are many different approaches available for analyzing such data including hierarchical linear models (HLMs), mixed models, fixed effects regression models, and multilevel structural equation models (MSEMs). Among these approaches, MSEM not only allows researchers to analyze multilevel data by investigating both within- and between-models simultaneously but also takes the potential measurement errors into account. Researchers can use the common structural equation modeling software (e.g., Lisrel, Mplus, and EQS) to conduct the multilevel covariance structural analysis. Nevertheless, very little effort has been made to assess the effectiveness of the traditional model fit indices on evaluating the misspecification in MSEMs. The purpose of this study was to investigate the sensitivity of the fit indices to the model misspecification under different conditional models. Factors including number of groups in the between model, group size, and ICC were considered. Our simulation results showed that RMSEA is more sensitive to the misspecifications in the within model than the between model and is only recommended for accessing the fit of the within model. On the other hand, SRMR-Within (SRMR-W) and SRMR-Between (SRMR-B) are sensitive to the misspecifications in within-model and between-model, respectively. Implications of the findings are discussed.

**Hwang, H.**

CS-Wednesday-am-s56

McGill University, Montreal, Canada  
(heungsun.hwang@mcgill.ca)

### Simultaneous Two-Way Clustering of Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) is useful for exploring the associations among multiple categorical variables. MCA has been combined with cluster analysis in a unified framework so as to take into account potential

## ABSTRACTS OF THE CONTRIBUTIONS

---

cluster-level heterogeneity of respondents in the data. The combined approach to MCA and cluster analysis provides a joint map of displaying variable category points and the centroids of clusters of respondents at the same time. This helps understand the relationships between variable categories and clusters of respondents, and in turn, describe the clusters. However, the joint map is often difficult to clearly interpret which variable categories are related to a cluster particularly when the number of variable categories is large. To overcome this limitation, a new method is proposed that integrates MCA into cluster analysis to identify common clusters of both respondents and variable categories. The proposed method provides a joint map of variable categories and cluster centroids. It also offers cluster memberships of variable categories as well as respondents. An empirical application is presented to demonstrate the effectiveness of the proposed method as compared to the extant combined approach.

**Ip, E.**

IS-Tuesday-am-s25

Wake Forest University School of Medicine, Winston-Salem, NC  
(eip@wfubmc.edu)

Transition Models for Longitudinal Data Analysis with Applications to Studies in Social and Behavioral Sciences

Motivated by applications across many disciplines, the hidden Markov model (HMM) has been extensively developed and refined to handle multi-process longitudinal and spatial data. We focus on transition models for multivariate discrete HMM. The methodological challenge is the potentially large number of possible transitions and the nondirectionality of the probabilities of transition. When there are 10 latent states, the number of all possible transitions between states under a first-order Markov assumption is 100. Assuming that there are 5 important predictors that may affect transitions and that transitions are time homogeneous, a comprehensive linear model for transitions, even without interaction terms, would involve close to 500 parameters. If one further allows transition models to be time inhomogeneous, the number of parameters would multiply further. Except in cases when the dimension of problem is small, parsimonious models need to be considered. However, the pattern of transitions originating from state 1 could be very different from that originating from state 10, making parsimonious models difficult to construct. We present several approaches to solve this problem and illustrate these approaches with applications to several large national data sets. This is joint work with Frank Rijmen, Peter Zhang, Alison Snow Jones, and Jack Rejeski.

**Iwama, N.,** Toyoda, H., Suzukawa, Y., Kubo, S., and Takeshita, M.,  
and Ikehara, K.

CS-Wednesday-am-s54

Waseda University, Japan (n.i.0119@moegi.waseda.jp)

The Paired Comparison Method Utilizing a Multi Sample Analysis in SEM for the Case of Many Objects

In this presentation, we propose a new paired comparison model utilizing a multi sample analysis in Structural Equation Modeling (SEM). At first, we review Bradley-Terry model, Scheffe model, its modified models and Scheffe model in a framework of SEM. Then, we explain our new paired comparison model and confirm the efficiency of this model through two experiments about a naming test for new products of mineral water. Through the two experiments, it was cleared that we can compare the targets which belong to different comparison groups. The feature makes it possible that we keep the number of participants or that of comparison per him/her small and provides some merits in practical scenes such as a marketing research. In addition, the utilization of SEM provides an advantage that we are able to estimate and interpret not only the preference order for all targets but also the variability of individual preference for each target.

**Jahng, S.**, and Wood, P.K.  
University of Missouri, Columbia, MO (sj4m3@mizzou.edu)

Poster-Tuesday-pm-s50

Variance of Successive Difference in Stationary Time Series with Autocorrelation: A Temporal Instability Parameter

Temporal Instability of a time series process has been of interest in psychological research, such as mood instability in affective disorder. Using Intensive longitudinal data (ILD), such as Ecological Momentary Assessments, direct assessment on temporal fluctuations is available. Although variance,  $\sigma^2$ , and autocorrelation,  $\rho(1)$ , within a single series are the parameters representing two key components of instability (variability and temporal dependency), those are not a stand-alone index of temporal instability. We propose the variance of successive difference,  $\delta^2$ , as a population parameter of temporal instability of a stationary series and derive it as a function of  $\sigma^2$  and  $\rho(1)$ :  $E[(x_{i+1}-x_i)-E(x_{i+1}-x_i)]^2=E(x_{i+1}-x_i)^2=2\sigma^2-2E(x_{i+1}x_i)=2\sigma^2(1-\rho(1))$ . Three estimators of  $\delta^2$  were proposed:

$$\hat{\delta}_1^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} \left[ (x_{i+1} - x_i) - \frac{1}{n-1} \sum_{i=1}^{n-1} (x_{i+1} - x_i) \right]^2, \quad \hat{\delta}_2^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2, \quad \hat{\delta}_3^1 = 2\hat{\sigma}^2(1 - \hat{\rho}(1)).$$

A simulation study was conducted to investigate biasedness and effectiveness of the three estimators with 202,500 data sets generated from three known stationary ARMA processes, i.e., AR(1), MA(1), and ARMA(1,1). Result supports superior performance of  $\hat{\delta}_2^2$ , a.k.a., Mean Square Successive Difference (MSSD), to the other estimators, although their performance are asymptotically equivalent, especially between  $\hat{\delta}_1^2$  and  $\hat{\delta}_2^2$ . Given that MSSD is a successful estimator of  $\delta^2$ , a generalized multilevel modeling with gamma distribution and log link is proposed to model temporal instability using ILD, where MSSD is treated as a random effect.

**Jansen, M.G.H.**  
University of Groningen, Groningen, The Netherlands (G.G.H.Jansen@rug.nl)

CS-Monday-am-s8

A Comparison of Latent Trait Models for Speed Tests with Different Distributional Assumptions

Response times are considered relevant in a wide variety of psychological and educational measurement situations. In particular the use of computers for test administration has opened the possibility to study observations that are not available in paper and pencil tests, such as the response times on individual items. Response times can be seen as (achievement) measures in their own right, but also as explanatory variables for other types of test behavior; in the context of survey research for instance, response times to attitude questions might be related to (in)stability of attitudes, faking etc. In this paper we focus on so-called speed tests. We consider a number of different approaches for modelling item- or test response latencies, among these the model assuming gamma distributed response times, with fixed test and subject parameters, known as the Rasch model. In the extended Rasch model a mixed regression model with explanatory variables on subject and test level is incorporated. Other distributions than the gamma distribution are suitable for modelling continuous positive random variates. For instance the lognormal distribution has also been used for modeling response times. We will provide some empirical examples where the data are analysed, using a number of different approaches, and the results are compared.

**Javaras, K.N.**, and Ripley, B.D.  
University of Wisconsin, Madison, WI (javaras@wisc.edu)

IL-Monday-pm-s15

Latent Variable Models for Likert Attitude Data

Likert attitude scales are widely used to measure individuals' attitudes toward a particular entity, an example being attitudes towards one's own nation ("national pride"). Individuals are presented with several attitudinal statements about the entity and asked to indicate their level of agreement with each statement by choosing from ordered response categories. The resulting responses reflect individuals' underlying attitudes, but also their response style,

## ABSTRACTS OF THE CONTRIBUTIONS

---

which is defined as a consistent and content-independent pattern of response category selection, such as a tendency to agree with all statements (“acquiescence”). In the real data example we use here, differences in British and American responses to a Likert scale assessing national pride may reflect national differences in acquiescence, as well as national differences in national pride. Failure to control for response style can confound inferences about attitudes. We introduce a new measurement model for Likert attitude data, an “unfolding” latent variable model that allows not only multiple underlying attitudes, but also response style, to affect responses. In simulation experiments where response style differs between individuals, our model yields less biased inferences about attitudes than do other models, including the popular Likert scoring method, especially when there are unequal proportions of favorable and unfavorable statements.

**Jöreskog, K.**

CS-Wednesday-am-s65

Norwegian School of Management, Oslo, Norway (karl.joreskog@dis.uu.se)

### Factor Analysis of Ordinal Data: FIML vs. DWLS

This paper concerns factor analysis of ordered categorical items (Likert Scales). Two approaches are compared. One is the full information maximum likelihood method which maximizes the full multivariate multinomial likelihood (Jöreskog & Moustaki, 2006); the other is a two-step procedure which first estimates polychoric correlations and then fits the factor loadings by Diagonally Weighted Least Squares (DWLS), also called Robust Weighted Least Squares (RWLS). The latter uses the asymptotic covariance matrix of the polychoric correlations to estimate the standard errors of the factor loadings. Results from two empirical data sets and one small simulation study are reported. These results suggest that, although standard errors are slightly smaller for FIML, differences in estimates of the factor loadings and their standard errors are so small that they are unimportant for most practical purposes.

Takane, Y., **Jung, K.**, and Hwang, H.

CS-Wednesday-am-s54

McGill University, Montreal, Canada (kwanghee.jung@mail.mcgill.ca)

### The Regularized Reduced-Rank Growth Curve Models (GMANOVA)

The reduced rank growth curve model (GMANOVA) is a useful technique for investigating patterns of change in repeated measurements over time. We incorporate a ridge type of regularization into the model in such a way that separate ridge parameters are allowed in column (representing subjects) and row (representing variables) regressions that are followed by the generalized singular value decomposition (GSVD) for rank reduction. Permutation tests are used to choose the best dimensionality in the solution, and the  $G$ -fold cross validation method is performed to choose an optimal value of ridge parameters. A bootstrap method is used to assess the reliability of parameter estimates. An illustrative example is given to illustrate the use of the proposed method. The proposed model is further extended to a mixture model of GMANOVA and MANOVA.

**Jung, Song**

Poster-Tuesday-pm-s50

Sung-Kyun-Kwan University, Seoul, Korea (djflqjfl1012@naver.com)

### Development of A Shortened Form of the MMPI-2 using Full-Information Item Factor Analysis Method

The Minnesota Multiphasic Personality Inventory(MMPI) is one of the most popular and widely used personality assessment instruments in various settings. The primary advantage of the MMPI is to provide an accurate and comprehensive personality profile based on affluent data. On the other hand, a shortcoming of the MMPI is that length of the test instrument and amount of time required for test administration are too long. Therefore, many researchers have developed short forms of MMPI, which were of less items and less time consuming than the full MMPI, while retaining its validity and clinical utility. The purpose of this study is to develop a short form of MMPI-2 using full-information item factor analysis(FIFA) methods. In this study, I will analyze only MMPI-2 Depression

## ABSTRACTS OF THE CONTRIBUTIONS

---

scale which contains 57 items. In addition to that, results of the three factor analysis methods will be compared : (1) FIFA method as implemented in TESTFACT (2) analysis with tetrachoric correlation as implemented in SAS (3) analysis with tetrachoric correlation as implemented in MicroFact.

Takane, Y., and **Jung, Sunho**  
McGill University, Montreal, Canada (sunho.jung@mail.mcgill.ca)

CS-Wednesday-am-s56

### Nonsymmetric Correspondence Analysis and Simpson's Paradox

Nonsymmetric correspondence analysis (NSCA) was designed to analyze a directional relationship between rows (predictive categories) and columns (criterion categories) of a two-way contingency table. A contingency table is often accompanied by some external information on rows of the table. Constrained NSCA has recently been proposed to incorporate such additional information into NSCA as linear constraints on predictor categories. Imposing the constraints on the predictor categories also means that part of the predictive relationship is left unaccounted for by the constraints. This paper proposes a method of analyzing the residual part along with the part accounted for by the constraints in constrained NSCA. This helps to gain some insight into what is known as Simpson's paradox in the analysis of contingency table. Permutation tests are used to determine the best dimensionality of the solution space, and test the significance of various effects of the rows of contingency table on the columns. A bootstrap method is performed to evaluate the stability of the solution space. An illustrative example is given to demonstrate the usefulness of the proposed procedure.

**Junker, B.W.**  
Carnegie Mellon University, Pittsburgh PA (brian@stat.cmu.edu)

SA-Wednesday-pm-s67

### Beyond MCMC

Markov Chain Monte Carlo (MCMC) techniques have simplified the development of model estimation and simulation methods, extending our reach as applied statisticians and abetting a revolution in model development and unification in psychometrics. For example, more than one in four recent papers in *Psychometrika* make use of MCMC techniques in areas such as Bayesian structural equation modeling, mixtures of vector and ideal point latent utility models for preference data, computation of the exact null distribution for frequentist hypothesis tests in Rasch and social network models, model evaluation and comparison in cognitive diagnosis modeling, item response models for guessing behavior, and dependence models for conjoint choice experiments. MCMC methods can be viewed as instances of successive substitution methods, reaching back to Gauss-Seidel; as instances of numerical integration methods; or as alternatives/extensions to E-M for dealing with data augmentation and other forms of missing data. In this talk I will place MCMC in the context of other optimization and integration methods, and survey some extensions, refinements and alternatives to MCMC. A central challenge for the future of psychometrics is posed by the enormous data gathering and data storage capacities of modern computing and communications technology. Although MCMC methods will always be a sensible early estimation choice while developing new parametric models, they do not always scale up well as the dimension or sample size of the data increases. I will point to some situations where MCMC does not scale well, and indicate some computational alternatives that may work better.

**Kahraman, N.**, and De Champlain, A.  
National Board of Medical Examiners, Philadelphia, PA  
(nkahraman@nbme.org)

CS-Wednesday-am-s58

### Assessing the Underlying Structure of Communication and Data Gathering Skills on a Sample of USMLE Step 2 CS Cases Using Confirmatory Factor Analysis

Standardized patients (SPs) are laypersons trained to portray clinical encounters in a consistent, standardized fashion; the latter simulated patient problems are referred to as cases and are organized in a clinical skills examination (CSE).

## ABSTRACTS OF THE CONTRIBUTIONS

---

During each case encounter, examinees are asked to engage in a variety of activities such as history taking, performing a physical examination, etc. The length of time spent at each station typically ranges from five to 20 minutes after which specific examinee skills are assessed by the SP via rating scales and/or observational checklists. As of June 2004, a clinical skills component was added to the United States Medical Licensing Examination (the USMLE Step 2 CS) and must be successfully completed by all candidates wishing to be licensed to practice medicine in the United States. The purpose of this study is to gather evidence of structural validity with regard to communication and data gathering (checklist) scores. More specifically, the aim is to assess whether the underlying structure of a given case response matrix differs as a function of the SP portraying the case. That is, is the underlying structure of a case response matrix comparable, irrespective of the SP portraying and rating that clinical scenario? This question will be addressed by initially fitting a confirmatory factor model that will treat each case generically, i.e., without any regard to the particular SP involved. In a second step, a model that takes into account the specific SP involved in the interaction with the examinee will be fit to the data matrix using a structurally missing case-SP design. It is hoped that these results will provide useful practical information not only with respect to scoring and psychometric issues, but also in regard to the training of standardized patients.

**Kan, K.J.**, van der Maas, H.L.J., and Dolan, C.V.

CS-Wednesday-am-s55

University of Amsterdam, Amsterdam, The Netherlands (k.j.kan@uva.nl)

### A Dynamical Model of General Intelligence: The Positive Manifold of Intelligence by Mutualism

Scores on cognitive tasks used in intelligence tests correlate positively with each other, i.e., they display a positive manifold of correlations. The positive manifold is often explained by positing a dominant latent variable, the g-factor, associated with a single quantitative cognitive or biological process or capacity. In this talk we discuss a new explanation of the positive manifold based on a dynamical model, in which reciprocal causation or mutualism plays a central role. It is shown that the positive manifold emerges purely by positive beneficial interactions between cognitive processes during development. A single underlying g-factor plays no role in the model. The model offers explanations of important findings in intelligence research. We will analyze the relation between factor models and mutualism models and discuss possibilities to test the models empirically.

**Kaplan, D.**

CS-Tuesday-pm-s47

University of Wisconsin, Madison, WI (dkaplan@education.wisc.edu)

### Statistical Considerations in a Counterfactual Theory of Causation for Non-Experimental Studies with Implications for Structural Equation Modeling

This presentation links a variant of the counterfactual theory of causation to the statistical problem of exogeneity. Specifically, I begin by arguing that Mackie's (1980) counterfactual theory of causation, extended by Hoover (2001), represents a sound philosophical basis for causal inference in non-experimental studies. Nevertheless, in the context of statistical practice, it is essential to examine the consequences of what it means to counterfactually vary the causal variable within a theoretically specified causal model. Here, Woodward's (2000) manipulability theory of causation plays a vital role. A key component of the manipulability theory of causation is the concept of invariance to interventions. A stronger form of invariance, referred to as *super-exogeneity* is discussed and linked to invariance. However, super-exogeneity, requires the statistical assumption of *weak exogeneity* to hold. Weak exogeneity addresses the problem of the ignorability of information in the marginal distribution of the causal variable as it pertains to the estimation of the conditional distribution of the outcome given the causal variable. In this talk, I argue that weak exogeneity is necessary but not sufficient to yield an unbiased estimate of the causal effect. A comparison of classical versus Bayesian structural modeling approaches to this problem is provided.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Karelitz, T.**<sup>(1)</sup>, and Budescu, D.V.<sup>(2)</sup>

CS-Wednesday-am-s56

<sup>(1)</sup> PACE Center, Tufts University, Somerville, MA (t.karelitz@tufts.edu)

<sup>(2)</sup> University of Illinois, Urbana Champaign

### Thinking Outside the Diagonal: An Exploratory Framework for Evaluating Agreement in Rectangular Contingency Matrices

Traditional indices of agreement, such as Cohen's kappa assume implicitly that raters use the same set of categories and the main diagonal of the contingency matrix is the primary source of evidence for agreement. Other methods, such as weighted or hierarchical kappa, use symmetric sets of off-diagonal cells in their calculation. We propose a more general framework that explores sequences of agreement patterns in co-occurrence matrices. These patterns are contiguous and monotonic, but may be asymmetric or exclude part of the diagonal. The proposed framework provides tools to analyze situations where traditional methods are inappropriate or their assumptions are violated. For example, it allows exploratory analysis of agreement in rectangular matrices, where the notions of main diagonal and symmetry are ill-defined. We use a dynamic algorithm to identify a hierarchical set of agreement patterns. At each step, the algorithm selects patterns that maximize different agreement or association indices, while preserving monotonicity. The algorithm provides estimates of agreement for each pattern and compares them to meaningful benchmarks. Researchers can use this analysis to investigate the appropriateness of the rating categories with respect to a given sample, or the extent to which judges agree, given the way they use the rating categories.

**Kateri, M.**<sup>(1)</sup>, Iliopoulos, G.<sup>(1)</sup>, and Ntzoufras, I.<sup>(2)</sup>

IS-Tuesday-pm-s43

<sup>(1)</sup> University of Piraeus, Piraeus, Greece (mkateri@unipi.gr)

<sup>(2)</sup> Athens University of Economics, Athens, Greece

### Bayesian Analysis of the Order Restricted RC Association Model

**Abstract:** Association models constitute a powerful tool for the analysis of contingency tables. They impose special structure on the underlying association between the classification variables, by assigning scores on their levels, which can be fixed or parametric. Under the general row-column (RC) association model for two-way tables, both row and column scores are unknown parameters without any restriction concerning their ordinality. In case of ordinal classification variables, order restrictions on the scores arise naturally, leading to an order restricted estimation problem, which is faced here Bayesian. The proposed procedure moves across models of different dimension, in order to infer for the equality of subsequent scores. This is achieved by developing a reversible jump Markov Chain Monte Carlo algorithm, which estimates accurately the posterior model probabilities. The algorithm can be used either for model comparison or for model averaging. Characteristic illustrations are provided and commented.

**Kato, K.**

Poster-Tuesday-pm-s50

University of Minnesota, Minneapolis, MN (kato0027@umn.edu)

### Improving Efficiency of Cognitive Diagnosis by Considering Incorrect Responses in Multiple-Choice Items: A Computerized Adaptive Testing Perspective

Several cognitively diagnostic psychometric models for nominal response data have been proposed recently in order to improve efficiency of cognitive diagnosis. This study focuses specifically on diagnosing students' use of defective strategies (cognitive rules), using a polytomous latent class model for the situation in which different incorrect responses are related to different cognitive rules in multiple-choice items. Previous studies indicate that diagnostic efficiency can be improved by considering differentiating incorrect responses with respect to several criteria such as the Kullback-Leibler (KL) information when compared to the case in which items are scored dichotomously. In this study, we examine the amount of improvement from a computerized adaptive testing perspective. Polytomous and dichotomous latent class models are fit to Siegler's Balance Scale data, which consist of multiple-choice responses of 719 students to 20 items. Diagnostic performance of adaptive testing is compared between the polytomous and dichotomous models. Items are selected by the global diagnostic index based on the KL information and the Shannon

## ABSTRACTS OF THE CONTRIBUTIONS

---

entropy, and the number of items required to reach a diagnosis at a certain level of accuracy is recorded. Preliminary results show that the number of required items is substantially reduced by considering incorrect responses.

**Kawahashi, I.**, and Toyoda, H.

Poster-Tuesday-pm-s50

Waseda University, Tokyo, Japan (ikko-kawa@aoni.waseda.jp)

### A Paired Comparison Model Including Individual Difference in Utility and Social Desirability

Due to an influence of social desirability, subject's response to personality test constructed by Likert scales are often biased. For resisting the bias, it is effective to apply a paired comparison scaling method represented by structural equation modeling (SEM). In this study, we attempt to construct a paired comparison EQ test which has a more robustness against the social desirability than a conventional scaling methods represented by a Liket method. Proposed model has several important features: (1) Based on Scheffe's paired comparison model, we add new parameters indicating individual difference in a utility and a social desirability of item to the model. (2) Including these new parameters in the model, we can evaluate a influence of social desirability to subject's response and remove it from scale scores(factor scores). As the result, we could confirm a high validity of proposed model as follows: (1)As compared with scale scores computed under Likert scales, new method had robustness against a social desirability. (2) From the view point of a latent mixture model, proposed model including parameter of individual difference in social desirability showed well fitting to data measured at recruit situation in which an influence of social desirability is non-ignorable.

**Kelderman, H.**

CS-Monday-pm-s19

Vrije Universiteit Amsterdam, Amsterdam, The Netherlands  
(h.kelderman@psy.vu.nl)

### Measuring the Nomothetic Meaningfulness of Operational Definitions of an Attribute

In multivariate behavioural research, theory is almost always formulated in terms of relations between subjects' attributes. If the relations of a certain attribute is much affected if we exchange one operationalization (test, item) for the attribute with another, the operational definition has little scientific value (Lazarsfeld, 1959). Suppes (1959, see also Narens, 2007) defined a measurement as *empirically meaningful* if the truth value of quantitative empirical statements based on it are invariant under arbitrary changes of the scale representation. Similarly, we consider their empirical meaningfulness with respect to replacements of operationalizations for a certain attribute within the nomological net (Cronbach & Meehl, 1955), called the *nomothetic meaningfulness* of the operational definition. In this paper we formulate this as an exchangeability property of the joint statistical distribution of the variables of interest (DeFinetti, 1937, see also Kelderman, 1995, 2004, 2008). We consider the equating of measurements from a domain towards nomothetic meaningfulness and linear statistical models that can be the used to study the nomothetic meaningfulness of equated scores. Statistics are proposed measuring the degree of nomothetic meaningfulness of a set of test items. The effects on nomothetic meaningfulness statistics is studied in a simulation and its use illustrated on empirical data.

**Keller, L.A.**<sup>(1)</sup>, and Keller, R.<sup>(2)</sup>

TS-Monday-pm-s22

<sup>(1)</sup>University of Massachusetts, Amherst, MA (lkeller@educ.umass.edu)

<sup>(2)</sup>Measured Progress, Dover, NH

### The Long-Term Sustainability of Different Scaling Methods in Item Response Theory True Score Equating

Item Response True Score equating requires that item and person parameters from different testing administrations be placed onto a common metric before equating. There are several popular methods for placing the parameters onto the same scale, and previous research regarding the relative performance of the methods indicates that the different

## ABSTRACTS OF THE CONTRIBUTIONS

---

methods do produce differences, sometimes substantial differences, when the distribution changes between the administrations. Given the results of the previous research, the quality of estimates that result from equating across multiple administrations may be in question, especially in cases where the ability distribution would be expected to change. However, there is no research that has investigated the accumulation of error for various scaling methods over multiple administrations. Therefore this study examines 6 popular scaling methods, over a six-administration chain of scaling. Two different anchor test lengths were considered, and three different series of ability distributions were simulated. Results are evaluated by examining the bias and root mean square error in the resulting parameter estimates as well as the classification accuracy in the case where examinees are classified into one of four performance categories. Results indicate that FCIP-1 did not perform acceptably in many cases, and FCIP-2, Haebara and Stocking & Lord performed the best, with mean-mean and mean-sigma in between these sets of results.

**Keller, R.**<sup>(1)</sup> and Keller, L.A.<sup>(2)</sup>

TS-Monday-pm-s22

<sup>(1)</sup> Measured Progress, Dover, NH (Keller.robert@measuredprogress.org)

<sup>(2)</sup> University of Massachusetts, Amherst, MA

### The Effect of Changing Scaling Methods in Item Response Theory True Score Equating

Several methods exist for placing the item and person parameters of item response theory onto a common scale. This is commonly done as a first step in IRT true score equating and previous research indicates that there are differences in the methods under certain conditions. Specifically, problems in biased estimation resulting from using certain fixed common item parameter (FCIP) methods have raised concerns, and Kim (2006) investigated different implementations of the FCIP method that can help improve the bias. However if the original implementation of FCIP (FCIP-1) has been used as a scaling method in an operational testing program, changing the scaling method would lead to different inferences being made about examinee performance as there would appear to be changes in performance that might be attributable to the reduction of bias from changing scaling methods. Therefore, this study seeks to understand the effect of changing scaling methods after the FCIP-1 method has been used for several administrations. In this study 5 popular scaling methods are investigated for a three-year administration period following three years of using FCIP-1. Results indicate that the greatest reduction in bias was found when using FCIP-2, Haebara or Stocking & Lord, and less dramatic results were found with mean-mean and mean-sigma.

**Kelley, K.**

IS-Monday-am-s5

University of Notre Dame, Notre Dame, IN (kkelley@nd.edu)

### The Average Rate of Change in Longitudinal Models

In the literature the slope from the straight-line change model has been interpreted as if it were equal to the average rate of change. It is shown, however, that this is generally not the case and is true in only a limited number of situations. The concept of the average rate of change is discussed in general as its precise mathematical definition. General equations for the discrepancy (raw and proportional) between the slope from the straight-line change model and the average rate of change are derived for discrete and continuous time models. Realizing that the slope from the straight-line change model is not generally equal to the average rate of change, methods for estimating the average rate of change are proposed. The potential benefits and drawbacks of supplementing an appropriate change model with an estimate of the average rate of change in applied research are considered.

**Kenny, D.A.**

KN-Wednesday-am-s51

University of Connecticut, CT (david.kenny@uconn.edu)

### Fixed and Random Effects Communicating to Each Other: Examples from Dyadic Research

Most models have two pieces: a fixed and a random piece. Normally in modeling little or no attention is played to the coordination of these two parts of the model. I suggest how these two different types of effects can be related

## ABSTRACTS OF THE CONTRIBUTIONS

---

using examples from my research area of dyadic analysis. First, I discuss how specification error in one part of the model affects the other part of the model. As an example, I use the measurement of consensus in person perception. Second, I discuss how a fixed effect is often closely tied to given random effect and I argue that an analysis of the random effects can be used to re-conceptualize a fixed effect. As an example, I consider the effect of group diversity on group cohesiveness. Third, I consider the relative power of tests of fixed and random effects and I argue that fixed effects often have considerably more power than random effects. As an example, I consider the measurement of gender differences in nonverbal sensitivity.

**Khalid, M.N.**, and Glas, C.A.W.

TS-Tuesday-am-s38

University of Twente, Enschede, The Netherlands (m.n.khalid@gw.utwente.nl)

### A Stepwise Method for Evaluation of Differential Item Functioning

The issue of item bias or differential item functioning (DIF) has an important impact on the field of psychological and educational measurement. In this paper, DIF is seen as a lack of fit to an item response (IRT) model. The method proposed here addresses the following problems. (1) Item-oriented test statistics are dependent, and if a number of items have DIF, this dependency may cause problems in identifying the true DIF items. (2) The power of the test statistics increases with the sample size. Therefore, it is not only significance but also effect size that determines whether an item should be flagged as having DIF. (3) However, the between-items dependence of the estimates of the effect sizes may also complicate the inferences. To address these problems, a model fit procedure based on the Rao's efficient test score is presented for the two parameter logistic model (2PL). The effect sizes were related to the consequences of DIF on pass/fail decisions and on inferences regarding parameters of the population distributions (i.e., differences in ability between populations). The procedure identifies items with DIF one at the time, models DIF by the introduction of group-specific item parameters and uses effect sizes as a stopping criterion. A simulation study was conducted to investigate the Type I error rate and Power of the test using different test lengths, sample sizes, numbers of biased items, and degree of misfit. The results showed that the procedure has a good Power to detect model violations and has a Type 1 Error rate that is acceptably close to the nominal significance level. Further, the simulation studies show that the dependence between the tests does have an important effect on the identification of DIF and that disregarding this dependence does lead to incorrect inferences.

**Kim, I.-H.**

Poster-Tuesday-pm-s50

Institute of Applied Psychology, Graduate School of Sung-Kyun-Kwan University,  
Seoul, South Korea (kb1121@naver.com)

### Problems of and Suggestions for Multi-Group Confirmatory Factor Analysis in Structural Equation Modeling

Using structural equation modelling (SEM), multi-group confirmatory factor analysis(CFA) is widely used for testing models. However, it has not been noticed that group effect in multi-group CFA is not focused. In the socio-cultural revolution of behavioral science since 1980's, situation has been recognized as a systematic factor that needs to be a part of theory construction. So, using situation factor as a notice group of common factors, instead of method in multitrait-multimethod (MTMM) analysis, would open an avenue of introducing socio-cultural perspective in modeling a covariance data set. For this modeling, we need a multitrait-multisituation (MTMS) approach. The purpose of this study is to introduce the process of multitrait-multisituation-multimethod (MTMSMM) approach which is a combined analysis of typical multi-group CFA and MTMS analysis. MTMSMM analysis is different from multitrait-multimethod-multitime (MTMMMT) analysis of Saris & Andrews(1991) conceptually. First, MTMMMT model is a combination repeated multimethod (RMM) and multi-trait approach. MTMSMM analysis is a combination of multi-group CFA and MTMS analysis. Second, MTMMMT model was introduced to investigate reliability and validity of measures. But MTMSMM model is proposed to pursue generalization of SEM involving measurement quality and structural relationships. In our study, MTMSMM approach is performed to demonstrate the utility of our approach. For example, the several tests for factorial invariance processes will be performed: configural

## ABSTRACTS OF THE CONTRIBUTIONS

---

invariance, metric invariance, scalar invariance, invariance of error variance, invariance of factor variance, factor mean invariance testing. Keywords: SEM, multi-group confirmatory factor analysis, MTMM, MTMSMM.

**Kim, J.S.**, Frees, E.W., and Swoboda, C.M.  
University of Wisconsin-Madison, Madison, WI (jeeseonkim@wisc.edu)

CS-Wednesday-am-s57

### Multilevel Model Specification Tests Using the Generalized Methods of Moment (GMM) Estimation Techniques

Standard multilevel model estimators such as ML and GLS assume that predictors are independent from random components. Although ML and GLS estimators are not even consistent without this assumption, orthogonality is not satisfied in many applications, and some predictors and random components are correlated. This condition is referred to as *correlated effects* or *endogeneity of predictors*. Correlated effects in multilevel models may yield severe bias in both regression coefficients and variance components at any level. By employing GMM estimation techniques and exploiting the hierarchical nature of multilevel data, this study presents a series of tests for examining the severity of correlated effects in multilevel models and also provides a battery of alternative estimators that are robust in the presence of correlated effects. It is shown that GMM approach provides an overarching framework that encompasses well-known estimators such as fixed and random effects estimators and also provides more options along a robust to efficient continuum. In addition, one-degree-of-freedom tests for each regression coefficient and the adaptation of empirical standard errors are demonstrated. The properties of multilevel model specification tests and the GMM estimators will be discussed in conjunction with a summary of simulation studies and findings from real data analysis.

**Kim, K.H.**, and Hsieh, J.-C.  
University of Pittsburgh, Pittsburgh, PA (khkim@pitt.edu)

TS-Tuesday-am-s31

### Power and Sample Size in Nested Covariance Structure Models

The relation among fit indexes, power, and sample in nested covariance structure models is examined. A power study for nested covariance structure models computes a minimum sample size required to achieve a certain power for constrain parameters between models not a power for entire model. In SEM, the power of a model as well as the power for specific parameter(s) is of interest. So far, a computation of power for a parameter has been difficult in SEM. In this study, 4 fit indexes (RMSEA, CFI, McDonald's Fit Index, and Steiger's gamma) were used to compute a minimum sample size. The resulting power and sample size depend on (1) number of parameters constrained, (2) change in model fit (i.e., difference in fit index between two models), (3) relation among the variables, and (4) value of fit index. A power study for a multi-sample SEM (e.g., measurement invariance, structural invariance) can easily performed using the proposed method.

**Kim, W.**<sup>(1)</sup>, Finkelman, M.<sup>(2)</sup>, and Nering, M.<sup>(1)</sup>  
<sup>(1)</sup> Measured Progress, Dover, NH (wkim@measuredprogress.org)  
<sup>(2)</sup> Harvard School of Public Health, Boston, MA

IS-Tuesday-am-s33

### A New Person-Fit Method for Short Tests

The purpose of person-fit analysis is to detect response patterns that are unlikely given a hypothesized test theory model or are aberrant compared with the majority of response patterns in the sample. The problem of empirical Type I error rates that deviate from the nominal level (e.g., 0.05) has been reported in person-fit analyses. In short tests, the empirical Type I error rate of person fit statistics may be too low, which is problematic because it will imply a reduction of statistical power. In this study, a new method is proposed to cope with such a problem. The method is based on the fact that all possible response patterns can be considered when the test length is small. Because of this fact, a rule can be created that yields increased power, yet maintains the nominal Type I error rate for all possible

## ABSTRACTS OF THE CONTRIBUTIONS

---

values of the latent trait. Using both simulated and real data, the new method is compared with (1) the parametric log-likelihood (Levine & Rubin, 1979) method, which uses an asymptotic normal distribution, and (2) the parametric person response function (PRF) (Trabin & Weiss, 1983) method, which uses an asymptotic chi-square distribution.

**Kim, Y. Y.**, and Reckase, M.D.

CS-Monday-pm-s19

NAEP-ESSI at American Institutes for Research, Washington, DC (ykim@air.org)

### A Psychometric Approach to Evaluating Constructs Comparability and its Application to Measure Construct Ability Growth

In test equating and/or linking approaches, comparability of constructs of different forms of test or similar tests has been assumed without being evaluated psychometrically. One possible reason is that no psychometric procedure has been developed yet. In this study, a multidimensional response theory (MIRT) approach to compare constructs comparability is outlined and how to measure growth of mathematical abilities at construct level is illustrated using real mathematics test data from a large-scale assessment program. Two tests were calibrated using BMIRT program (Yao, 2003) and the results were linked through fixed common item parameters method using BMIRTanchor program (Yao, 2003). Then, each set of items with similar combination of abilities/skills were identified through MIRT cluster analysis and the direction of each reference composite as the best measurement of the combination of abilities/skills for each set of items was identified in the ability space specified by the tests. The directions of each reference composite were compared to evaluate constructs comparability. Identified reference composites were aligned with coordinate axes in a new coordinate system in the ability space applying procedures outlined by Reckase (in press). Then, constructs for each test were measured and growth on each construct was evaluated.

**Kingsbury, G.G.**, Hauser, C., and Wise, S.L.

IS-Monday-pm-s18

Northwest Evaluation Association, Lake Oswego, OR (gage.kingsbury@nwea.org)

### The Impact of Individual Validity on Item Calibration

Recently, Kingsbury and Hauser (2006) introduced the concept of individual validity as a structure for quantifying whether, and to what extent, a particular test score is a reasonable representation of the capabilities of any particular test taker. It is clear that if an individual is distracted, disengaged, or cheating on a test, the test score obtained will be less useful for making decisions or for measuring change. It is somewhat less clear what the effect of low individual validity might be on the development and maintenance of Item Response Theory (Lord and Novick, 1968) measurement scales. The current study investigates how differing levels of individual validity might influence item difficulty estimates obtained from the 1-parameter logistic model. The first portion of the study uses simulation to study the change in item parameter estimates as individual validity decreases. The second portion of the study applies the individual validity procedures suggested by Kingsbury, Hauser, and Wise (2008) to a live data set. In this portion of the study, the procedures are used as a filter for parameter estimation and the differences in item difficulties with and without filtration are examined.

van der Linden, W.J. **Klein Entink, R.**, and Fox, J.P.

CS-Monday-am-s8

University of Twente, Enschede, The Netherlands (r.h.kleinentink@gw.utwente.nl)

### IRT Parameter Estimation with Response Times as Collateral Information

Hierarchical modeling of responses and response times on test items facilitates the use of response times as collateral information in the estimation of the response parameters. Two sources of collateral information are identified: (i) the joint information in the responses and the response times summarized in the estimates of the hyperparameters and (ii) the information in the posterior predictive distribution of the response parameters given the response times. The latter is shown to be a natural empirical prior distribution for the estimation of the response parameters. Unlike traditional hierarchical IRT modeling, where the gain in estimation accuracy is typically paid for by an increase in bias, use of

## ABSTRACTS OF THE CONTRIBUTIONS

---

this posterior predictive distribution improves both the accuracy and the bias of IRT parameter estimates. In an empirical study, the improvements are demonstrated for the estimation of the person and item parameters in the 3-parameter response model.

**Konya, Y.**, Shimizu, Y., and Kano, Y.

CS-Wednesday-am-s64

Osaka University, Osaka, Japan (konya@sigmath.es.osaka-u.ac.jp)

### Interval Estimations Based on Normalizing Transformations by Two Approaches

Although the sampling distribution of a statistic plays an essential role in interval estimation, the exact one is generally unknown. When a Studentized statistic follows according to the normal distribution asymptotically, the normal approximation to the unknown sampling distribution is often used. However confidence intervals based on the approximation have large coverage error. Normalizing transformation is useful in reducing the error. Although the transformation could be obtained from the Edgeworth expansion for the Studentized statistic, the resultant confidence interval could be disconnected. We propose two approaches to this problem. They are called "monotonization" method and "selection" method. The former is suggested by Hall (1992), while the latter is entirely new. Their coverage errors are the same order as that of a BCa interval (Efron, 1987), and the coverage error of the BCa interval is said to be small enough in practical use. Numerical experiments are conducted to compare the performance of these methods and some existing ones including BCa intervals, which suggests that the proposed methods are the best and that the "selection" method has smaller coverage error than the "monotonization" method for small and medium sample cases.

**Kroonenberg, P.M.**<sup>(1)</sup>, Büyükkurt, B.K.<sup>(2)</sup>, and Mesman, J.<sup>(3)</sup>

IS-Tuesday-am-s26

<sup>(1)</sup> Leiden University, Leiden, The Netherlands (kroonenb@fsw.leidenuniv.nl)

<sup>(2)</sup> Concordia University, Montreal, Canada

<sup>(3)</sup> Leiden University, Leiden, The Netherlands

### Longitudinal Assessment of Child Behaviour: Comparing Models for Three-Way Binary Data

Child behaviour is often assessed with the more than 100 items of the Child Behavior Check List (CBCL). When such assessment is carried out more than once, a three-way data set is created of children by behaviours by measurement periods. Typically, extreme behaviour is rare so that the three-point rating scales are often very positively skewed. For the present model comparison the data are dichotomised into 0-1 data: The behaviour is present or absent. For such binary data the Tucker3-hierarchical classes (HICLAS) analysis is an appropriate tool to examine the patterns of behaviour and who has which pattern over time. The primary objective of this research is to compare a Tucker3-HICLAS analysis with the results of applying a Tucker3 model for three-way rating data to the binary data. The differences and similarities between the results of the two models will be examined. A further comparison can be made with Tucker3 analyses based on the (uncondensed) three-point rating scales. Finally, graphical displays that have been developed for the Tucker3 model (such as paired component plots, three dimensional plots, and joint biplots) will be compared with graphical displays developed for Tucker3-HICLAS analysis. The aim is to examine to what extent the models are complimentary or rivals for the same information.

**Kuha, J.**

IS-Monday-am-s6

London School of Economics, London, United Kingdom (j.kuha@lse.ac.uk)

### Missingness Factors in Latent Variable Modeling

When multi-item scales are employed in survey research, it is common that, for some respondents, responses are only obtained for some of the items. One approach to analysing the data in this incomplete-data situation is to consider latent-variable models which involve, in addition to substantive latent variables, also ones which are related solely to the propensity to respond to an item. The specification and estimation of such models are considered in this

## ABSTRACTS OF THE CONTRIBUTIONS

---

talk. The interpretation of the models is discussed, especially with reference to conventional classifications of missing-data mechanisms. The ideas are illustrated through the analysis of data on attitudes towards London's Metropolitan Police Force.

**Kunina, O.**<sup>(1)</sup>, Rupp, A.A.<sup>(2)</sup>, and Wilhelm, O.<sup>(1)</sup>

CS-Monday-am-s7

<sup>(1)</sup> Humboldt-Universität zu Berlin, Berlin, Germany

(olga.kunina@iqb.hu-berlin.de)

<sup>(2)</sup> University of Maryland, MD

### Convergence of Skill Profiles for Cognitive Diagnosis Models and Other Multidimensional Scaling Approaches: An Empirical Illustration with a Diagnostic Mathematics Assessment

A variety of multiple classification models, so-called *cognitive diagnosis models (CDMs)* (e.g., diBello, Roussos, & Stout, 2007; Rupp & Templin, 2007), have recently been developed and refined to respond to the demand for more fine-grained diagnostic inferences for formative educational purposes. However, the number of successful practical applications of these models is relatively small due to the high demands placed on assessment design, the large required sample sizes for these models, and the fact that only a limited number of moderately correlated latent attributes can be reliably discriminated in practice. In this paper, we apply a variety of compensatory and non-compensatory CDMs to pilot data from a diagnostic mathematics assessment in elementary school and compare them in terms of absolute and relative model fit, attribute difficulty distributions, and latent class membership probabilities. Furthermore, the discrete attribute profiles from the best-fitting CDM will be cross-validated with continuous ability profiles from multidimensional item response theory (e.g., Embretson & Reise, 2000) and confirmatory factor analysis (e.g., McDonald, 1999). In combination, these analyses will provide empirical insight into the conditions under which the theoretical potential of multiple classification models can be realized in educational assessment practice.

**Kupzyk, K.A.**

Poster-Tuesday-pm-s50

University of Nebraska, Lincoln, NE (kevink@bigred.unl.edu)

### A Comparative Study of Traditional and Bootstrap Methods for Tests of Mediation

Several studies investigating the tests of mediation effects have reported that the Sobel test for indirect effects suffers from low statistical power. This is due to the lack of normality in the sampling distribution of the product of the two direct effects in a mediation analysis. A few authors have suggested that the bootstrap standard error may be used as a proxy for the Sobel standard error in a normal-theory confidence interval. A simulation study was carried out to investigate the extent to which the bootstrap standard error diverges from the Sobel standard error. Results indicate that the Sobel and bootstrap standard errors were similar in magnitude as sample size grew large. At low sample sizes, Sobel standard errors were typically smaller. This indicates that the bootstrap standard error has lower power at low sample sizes if used in a normal-theory confidence interval. Conclusions should be based on the bootstrap confidence interval, which was found to have greater power, as others have reported. This study extended previous research by including a comparison of the bootstrap interval, the bias-corrected bootstrap, and the bias-corrected and accelerated bootstrap. Although the bootstrap interval was found to be smaller, the bias-corrected intervals had higher power.

**Kyungtae, K.**

Poster-Tuesday-pm-s50

Sung-Kyun-Kwan University, Seoul, Korea (kimkt21@naver.com)

### A Monte Carlo Study of Parameter Estimation in IRT

Two assumptions of local independence and unidimensionality are prerequisite for estimation of parameters in Item Response Theory (IRT). However, in practice these assumptions are not likely to be satisfied. In this research, by

## ABSTRACTS OF THE CONTRIBUTIONS

---

simulation study, I will test acceptable of degree of against robust assumptions. I generated data of IRT model under the favor of WINGEN2.0 (Han & Hambleton, 2007) simulation program, and analyze data set with Structure Equation Model(SEM) software. The conventional IRT with unidimensionality is regarded as a single factor model in SEM. Local independence in IRT is regarded as specification of measurement errors uncorrelated in SEM. The present study I generate data set of measurement errors correlated and analyze with IRT model program.

**Lane, S.P.**, and Shrout, P.E.

Poster-Tuesday-pm-s50

New York University, New York, NY (spl249@nyu.edu)

### Using Dynamic Factor Analysis to Assess Within-Person Reliability in Longitudinal Designs

Daily diary methods require brief assessments, and these may be more subject to measurement error than longer assessments. Cranford et al (2006) provided a method for estimating reliability of within-person assessments based on Generalizability Theory (GT) for measures with two or more fixed items. However, the GT framework assumes that items are parallel and it ignores the temporal dependence of measurement occasions. We provide a new method for estimating reliability in this context, using Dynamic Factor Analysis (DFA) and an adaptation of McDonald's (1999) omega reliability coefficient. The method is illustrated using a sample of 98 individuals who were preparing for the NYS bar examination and who completed five different mood scales twice a day for 44 days. We report the DFA results and resulting within-person reliability estimates for the anxiety and depression scales. The estimates were not much different from the GT estimates using the Cranford approach, suggesting that in this case the GT assumptions were appropriate.

**Lawrence, D.R.**

CS-Monday-pm-s20

Rochester Institute of Technology, Rochester, NY (drleas@rit.edu)

### A Forced-Classification Analysis of Paired-Comparison Data Subject to a Polychotomous Criterion Item

Forced classification, a procedure of dual scaling that enables the investigator to emphasize one or more items of interest, has been applied in the analysis of various types of data, including both incidence data and dominance data. Recent research in this area has focused on dominance data with the items to be emphasized "external" to the original data set. In fact, these so-called "criterion" items were constructed from a dichotomous incidence item—which in practice might well be a demographic variable of some sort—and then "merged" with the dominance data. In a more recent study that involved rank-order data, the criterion items were constructed from a single polychotomous incidence item. This present study offers a general procedure for including such a polychotomous incidence item in a forced-classification analysis of paired-comparison data—analyses of which have, thus far, been limited to the simpler dichotomous incidence item.

**Lee, J.**<sup>(1)</sup>, and Han, K.T.<sup>(2)</sup>

Poster-Tuesday-pm-s50

<sup>(1)</sup> Sung-Kyun-Kwan University, Seoul, Korea (happyjhlee@gmail.com)

<sup>(2)</sup> Graduate Management Admission Council

### Optimizing the Number of Items for each Module in Multi-Stage Testing

Multi-Stage Testing (MST) is a type of computerized adaptive test (CAT) that administers blocks of items to an examinee based on her/his provisional proficiency estimate (in fact, CAT can be seen as a special case of MST where each module has only one item). MST provides the appealing features of CAT (e.g., efficiency) while minimizing the problem related to item bank exposure. However, it is important to consider how many items comprise a block while maintaining the advantages of CAT. In this study, the optimal number of items for each block, which minimizes the standard error of proficiency estimation and ensures the accuracy of proficiency estimates, constraining the test length will be investigated via a series of Monte-Carlo simulation studies. Various ratios of the number of stages and the number of modules in each stage given number of items in each module will

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

be examined. The results of this study will provide test developers with a comprehensive illustration of MST and guidelines for constructing MST structures.

**Lee, M. D.**

IL-Tuesday-pm-s40

University of California, Irvine, CA (mdlee@uci.edu)

### Bayesian Graphical Modeling in Cognitive Science

Graphical modeling provides an easily understood and powerful framework for implementing Bayesian analyses in the cognitive sciences. Graphical models specify a probabilistic generative process for observed data in terms of psychologically meaningful parameters, and allow Bayesian inferences about parameters, data, and models to be made using computational sampling methods. We give a series of tutorial examples, trying to highlight those features of the Bayesian graphical modeling approach most likely to foster theoretical progress in understanding human cognition. These examples include models of stimulus representation, memory retention, and heuristic decision-making. We also give a brief survey of more advanced recent applications, spanning a range of topics in higher-order cognition, to try and indicate the generality and potential of the graphical modeling approach.

**Lee, S.**

TS-Tuesday-am-s31

Sung-Kyun-Kwan University, Seoul, Republic of Korea (smlyhl@chol.com)

### Choice Between a Common Factor Approach and a Correlated Uniqueness Model to Specify Method Effects in a Confirmatory Factor Analysis of MTMM Data

In analyses of multitrait multimethod data it has been an issue how to handle method effects without encountering serious problems in estimation (i.e., nonconvergence, improper solutions). Although a common factor approach (CFA) seems elegant for interpretation, it has suffered the problems of estimation. As an alternative a correlated uniqueness model (CU model) has been proposed and employed often. However, a CU model has not been free from critiques. First, a CU model does not allow to directly estimate method variances. Second and more critical issue is that systematic method or situation effect may be a valid factor having effect on outcome variables, instead of errors of measurement. From a socio-cultural perspective it is argued that systematic method or situation effects should not be specified as correlated uniqueness. In this study, it will be demonstrated that the parameters in a CU model may be expressed as functions of parameters in a CFA approach at certain conditions, leading to equivalence between a CFA approach and a CU model. Then the argument for a choice between the two would be fruitless.

**Lee, W.-C., and Kim, S.**

TS-Wednesday-am-s59

University of Iowa, Iowa City, IA (won-chan-lee@uiowa.edu)

### Information Functions for Transformed Scores

The primary purpose of the present paper is to describe and illustrate procedures for computing test and item information functions for scale scores that are transformed from either number-correct scores or IRT proficiency scores. Test information functions for various types of transformed scale scores including number-correct scores are compared. Item information functions are computed for IRT proficiency scores and transformed scale scores, and results are compared when constructing a test to meet a specific target test information function.

**Lee, Y.-H., and Kwok, O.-M.**

Poster-Tuesday-pm-s50

Texas A&M University, College Station, TX (jasviwl@neo.tamu.edu)

### An Alternative Way to Test Interaction Effects Between Dummy Variables

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

Researchers in social and behavioral sciences make extensive use of dummy coded variables for statistical analyses. Dummy variables have the advantage of simplicity in use and in interpretation. However, the interpretation of dummy variables may not be as straight forward as we generally expect when an interaction effect involving two dummy variables is included. In this paper we demonstrate that creating respective dummy variables instead of employing an interaction term between two dummy variables in a regression analysis yields the same result as in the ANOVA approach on testing interaction effect between two factors. The advantages of the former technique (i.e., creating respective dummy variables) include: 1) it can test the null hypothesis in one step without performing post hoc tests when an interaction term is statistically significant, and 2) it enables an easy interpretation of the significance test result. Limitations on utilizing such a coding scheme and other potential coding methods are provided.

**Lee, Y.-H.**, and von Davier, A.A.

IS-Tuesday-am-s32

Educational Testing Service, Princeton, NJ (ylee@ets.org)

### Alternative Kernels in the Kernel Equating Framework: Logistic and Uniform

The kernel equating method proposed by von Davier, Holland, and Thayer is based on a flexible family of equating functions with the Gaussian kernel. While the classical equipercentile equating method continuizes the discrete score distributions by linear interpolation, in principle the kernel equating method can smooth them with the use of various kernel functions. The logistic kernel shares much common ground with the Gaussian kernel with the exception of heavier tails. On the other hand, the continuous uniform kernel has no tail and is known to work as the linear interpolation. Continuous distributions produced by these kernel functions may resemble the discrete score distributions more closely than those produced by the Gaussian kernel for tests with different properties (for instance, easy or difficult). To illustrate, an application is considered for an equivalent-groups design.

**Lei, M.**<sup>(1)</sup>, and Lee, W.-C.<sup>(2)</sup>

TS-Wednesday-am-s66

<sup>(1)</sup> The College Board, New York, NY (mlei@collegeboard.org)

<sup>(2)</sup> University of Iowa, Iowa City, IA

### An Alternative Method for IRT Classification Accuracy Estimation

Classification accuracy is defined as consistent classification between observed scores and true scores (e.g., Huynh, 1976; Livingstone & Lewis, 1995; Rudner, 2001). This study first proposes an alternative IRT method for computing classification accuracy index where examinee ability distribution is estimated using the estimated IRT observed score distribution and test characteristic curve, and true score distribution is assumed either normal distribution or 4-parameter beta distribution. Their parameters are estimated using model based observed scores. The estimation of model based observed score distribution is going to be improved using a Bayesian procedure. Second, the study will compare the Rudner method (Rudner 2000) and the proposed IRT method. The data used in the study will be empirical scores from a large long existing testing program. Its classification accuracy has been established through the repeated administrations of the parallel forms to the relatively stable samples, which will be used as the criteria in this study.

**Leighton, J.P.**, and Gierl, M.J.

IL-Monday-pm-s14

University of Alberta, Edmonton, Canada (jacqueline.leighton@ualberta)

### Cognitive Diagnostic Assessment for Education: Theory and Applications

Billions of public dollars are spent every year in the United States, Canada, and abroad on large-scale student testing. Federal legislation such as the No Child Left Behind (NCLB) Act in the US and international testing programs such as the Organization for Economic Co-operation and Developments Programme for International Student Assessment (PISA) and the National Center for Education Statistics Trends in International Mathematics and Science Studies

## ABSTRACTS OF THE CONTRIBUTIONS

---

(TIMSS) reflect increasing expectations and demands for accountability by means of educational testing. While unidimensional summative scores have generally been useful for making comparisons among students, states, provinces, and even countries, this information has been less useful for revealing students academic strengths and weaknesses, and helping teachers and administrators improve learning and instruction. Consequently, there is a call for educational tests in science, math, and reading that can be used not only to evaluate students overall proficiency but also to identify students' cognitive strengths and weaknesses. This kind of *cognitive diagnostic* information could be used to rank students on a summative scale as well as to directly facilitate students learning and teachers instructional practices. The philosophy and practice of a newly emerging form of assessment, Cognitive Diagnostic Assessment (CDA), is described in this presentation.

**Li, D.**

IS-Wednesday-am-s60

Educational Testing Service, Princeton, NJ (dli@ets.org)

### Random Effects Models for Large Scale Assessment

Hierarchical latent regression models (HLRM) by Li & Oranje (2006) are extended to more general two-level random effects models (REM), in which both cluster and individual level covariates are involved. A special case of the REM, a random intercept model (RIM) is explored in details using NAEP 2005 reading data for grade 8 assessment. Plausible values are imputed by one cluster at a time instead of one student at a time based on the assumption that students within a cluster are correlated. The results from HLRM are compared with random intercept models as well as the fixed effects models (FEM) that are currently used for the operational analysis. The results show that the regression effects in the FEM are larger than the counterparts in HLRM, with the regression effects from RIM falling between HLRM and FEM. The standard errors for the regression effects are in the opposite order: the largest ones are found in HLRM, and the smallest ones in FEM. The estimation of the mean scale scores for subgroups (e.g., gender and race/ethnicity groups) using the HLRM, RIM, and FEM shows slight differences among each other, and the differences of the standard errors for the mean scale scores among the three models are also trivial.

**Li, F., and Cohen, A.S.**

TS-Tuesday-am-s38

University of Georgia, Athens, GA (feiming@uga.edu)

### A Modified Higher-Order DINA Model for Detecting Differential Item Functioning and Differential Attribute Functioning

This study will present a modified higher-order DINA model (de la Torre & Douglas, 2004) for use in detecting both differential attribute functioning (DAF) and differential item functioning (DIF). In this model, the source of construct-relevant (i.e., benign) DIF can be simultaneously separated from the construct-irrelevant (i.e., adverse) DIF. DIF detection ensures test fairness and improves test validity, whereas DAF detection provides a better understanding of differential knowledge structures across groups, by specifically identifying group strengths and weaknesses in terms of a set of cognitive attributes after conditioning on general ability. A Markov Chain Monte Carlo (MCMC) algorithm employing Gibbs sampling will be implemented to estimate all parameters in the model, including the estimates for the indices of DAF and DIF. Thus, the empirical posterior distributions for the parameters accounting for the differences can be directly used to check the existence of the DIF and the DAF. A simulation study will be done for a Type I error and power analysis under conditions, simulated to represent practical testing contexts, by manipulating sample size, ability distribution difference, Q-matrix complexity, attribute discrimination parameters, and different DIF and DAF combinations. Finally, a real-data example will be presented and interpreted.

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

**Li, L.,** and Bentler, P.  
UCLA Integrated Substance Abuse Programs, Los Angeles, CA  
(lilibo@ucla.edu)

CS-Tuesday-pm-s44

### Cutoff Criteria for Comparing Closely Matched Structural Equation Models

The traditional model comparison procedure selects nested structured models by evaluating the feasibility of equality constraints that differentiate the models. However, this approach is questionable in practice because the less restricted model can capture any additional, even minor and less interpretable, characteristics in the sample. For overcoming this problem and for the justification and preservation of the parsimony or interpretability of the restricted model, we propose instead to evaluate model close match, using the distance between two models, either as important supplementary information or as a criterion for nested model comparison. Based on MacCallum, Browne and Cai (2006), we develop a reasonable cutoff criteria to determine the degree of close match between models and illustrate it with two application examples.

**Li, Z.,** and Anderson, C.  
University of Illinois, Urbana-Champaign, IL (zli10@uiuc.edu)

IS-Tuesday-pm-s43

### Log-Linear Models as Rasch Models with Collateral Information

This talk presents log-linear by linear association (LLLA) models as Rasch models with the addition of collateral information. Both item covariates and person covariates are incorporated into the LLLA model. Pseudolikelihood estimation is used to estimate all the model parameters. The efficiency of estimates and standard errors are discussed. Collateral information may be used to make inferences about latent trait more precise. The proposed method successfully recovers the parameters, yields similar results as other approaches, and is computationally faster than SAS/NLMIXED, which uses Gaussian quadrature to obtain maximum likelihood estimate. One of the advantages of this approach is that it is straightforward to generalize it to multidimensional models.

**Ligtvoet, R.,** van der Ark, L.A., and Sijtsma, K.  
Tilburg University, Tilburg, The Netherlands (R.Ligtvoet@uvt.nl)

CS-Tuesday-am-s29

### A Coarse Approach to IRT for the Evaluation of an IIO

An invariant item ordering (IIO) means that the same ordering of items, according to their attractiveness or difficulty, applies to all subjects. An IIO facilitating the interpretation and comparability of test results. The evaluation of an IIO is often difficult due to the large sample sizes needed to reliably establish an IIO, and by the fact that those subjects are needed in the sample that are often the most sparse in the population. A coarse approach that aims to remedy these drawbacks is suggested for the evaluation of an IIO. This approach consists of identifying clusters of items which are invariantly ordered across a small number of ordered latent classes. The approach results in a loss of precision with respect to the ordering of individual items and also results in a less refined scale of the latent variable. However, using this approach we expect to gain a powerful tool that allows a reliable evaluation of an IIO for realistic sample sizes. The advantages and disadvantages of the coarse approach to the evaluation of an IIO are discussed and some preliminary results of applying the coarse approach to test data are presented.

**Lin, A.**  
Pearson, San Antonio, TX (anli.lin@pearson.com)

TS-Monday-pm-s22

### Comparing Kernel Equating and IRT Equating

## ABSTRACTS OF THE CONTRIBUTIONS

---

The purpose of this study is to look at the relationship and difference between Kernel equating and IRT equating. IRT is a parameter method of model fit. Kernel equating is a non parameter method of model fit. In Kernel equating, the power moments of model fit could be changed from low to high. Also, the Kernel equating was equivalent to the percentile equating in a small bandwidth and the linear equating in a large bandwidth. In the study, when the rank of the fitting moments and the bandwidth in Kernel equating were adjusted, main effects and interactions of two factors in the score difference between Kernel and IRT equating were displayed. In some rank of the fitting moments and the bandwidth, the effect of the Kernel equating is close to that of IRT equating. The difference and relationship are explained from the essentials of log-linear model fit and IRT model fit.

**Lin, H.**, Chang, H.-H., and Chang, C.-H.  
University of Illinois at Urbana-Champaign, Urbana, IL  
(lin34@uiuc.edu or haiyanlin06@gmail.com)

TS-Monday-am-s11

### Improving Item Pool Usage and Content-Balancing in the Measurement of Geriatric Depression

Computerized Adaptive Testing (CAT) of patient-reported outcomes (PRO) has become increasingly important in health outcomes research and practice. CAT-PRO has tried to employ and implement many well-developed approaches in educational testing. Among them, the most commonly used item selection method, i.e., maximum information method, is widely proposed to be adopted in CAT-PRO applications. However, using maximum information throughout the item selection procedure may lead to overexposure of the high  $a$ -parameter items while some others never being selected. This may cause inaccurate estimations of latent trait being measured when the assessment length is relatively short, which is common in PRO applications. As reported by Ware et al. (2003) in a simulation study about headache impact, 20 out of 54 items are functional while the remaining 34 items have never been utilized. Consequently, the measurement becomes incomplete and unreliable because patients might remember their patterns for responding and not concentrate on the process when asked the same questions repeatedly in routine assessments. This study proposes to adopt a two-phase content-balancing method which is incorporated with the STR\_C approach with the  $a$ -stratified method and content-blocking design. This method controls the item exposure rate and satisfies the practical constraints, content-balancing, without compromising measurement precision. In addition, it takes a flexible content-balancing approach, in which the number of items from each content area is restricted within a range rather than a fixed number. It is more appropriate for assembling individualized tests with various test length contingent on patients' answers and situations in PRO assessment. The simulation study will be conducted by using an existing item bank of geriatric depression to understand the improvement of the proposed method in clinical setting.

**Lin, N.**, and Woods, C.  
Washington University, St. Louis, MO (nlin@math.wustl.edu)

IS-Wednesday-am-s53

### Davidian-Curve Item Response Theory and F-Information of Item Parameters

We developed a nonparametric item response theory framework, Davidian-curve IRT (DC-IRT), using "seminonparametric" (SNP) representation of the latent density. Item parameters can be estimated using the marginal maximum likelihood based on the expectation-maximization algorithm. The SNP representation is highly flexible and only needs a few parameters for most applications. Due to its succinct form, we are also able to derive the F-information of the item parameters, which can be viewed as a generalization of the Fisher information to measure the information in a distribution with respect to only part of the unknown parameters. Assuming asymptotic normality of the marginal MLE, standard errors of the item parameter estimates can be obtained using the F-information.

**Linting, M.**

DP-Tuesday-pm-s49

Leiden University, The Netherlands (linting@fsw.leidenuniv.nl)

### Nonparametric Inference in Nonlinear Principal Components Analysis: Exploration and Beyond

This presentation will focus on nonlinear principal components analysis (NLPCA), a nonlinear alternative to standard linear PCA, used to explore the correlational structure among different types of variables (nominal, ordinal, and numeric) that may be nonlinearly related to each other. As PCA does not make distributional assumptions, it is theoretically insensible to apply standard (asymptotic) inferential formulas. In addition, such asymptotic methods have been shown to be less effective than nonparametric alternatives (Timmerman, Kiers, and Smilde, 2007). Therefore, the main purpose of this presentation is to show easily applicable methods for performing nonparametric inference on the elements of the NLPCA solution. First, the stability of the eigenvalues, component loadings, component scores, and category quantifications is established by using the nonparametric balanced bootstrap procedure (Efron, 1982; Efron and Tibshirani, 1993). Specific attention is paid to the graphical display of the results, and we show how to interpret and decrease instability. Second, the statistical significance of the NLPCA results is assessed, comparing two different permutation strategies in a simulation study in linear PCA: (1) independent and concurrent permutation of all variables in a data set (see Buja and Eyuboglu, 1992), and (2) permutation of one variable at a time, while keeping the other variables fixed. Method (2) is found the most effective for establishing the significance of the contribution of separate variables, and this method is applied to the NLPCA solution. Throughout the presentation, an empirical data set is analyzed as an illustration.

**Liu, J.**

TS-Monday-am-s11

Educational Testing Service, Princeton, NJ (jxliu@ets.org)

### Multidimensional Computer Adaptive Strategies in Psychological and Health Assessment

Given the popularity of computer adaptive testing (CAT) in achievement tests, some researchers suggest that CAT should be used in psychological and health assessment. These tests are often designed to provide comprehensive information along several dimensions of knowledge, attitude, or personality. This study aims to compare the efficiencies of multi-dimensional CAT versus uni-dimensional CAT based on the multi-dimensional graded response model. Item selection and ability estimation methods based on multi-dimensional graded response models will be developed. A real data CAT simulation will also be conducted. A dataset consisting of approximately 3,000 item responses to the DASH (Disabilities of the Arm, Shoulder, and Hand) and SF-36 (MOS 36-Item Short-Form Health Survey) from outpatients treated at Centers for Rehabilitation Services will be used. The item parameters from the DASH and SF-36 create a three-dimension item pool. The empirically based correlations between the three dimensions will be used in this study. The performance of the CAT is evaluated for fixed test lengths and different ability levels. The outcome measures include the correlation between estimated and true ability, root mean squared error, bias, and standard error for trait estimates.

**Loye, N.**

TS-Tuesday-pm-s48

University of Montreal, Montreal, Canada (nathalie.loye@umontreal.ca)

### Taking up the Challenge to use a CDM to improve Q-matrices: An Illustration

This presentation is an illustration of using the Cognitive Diagnostic Model (CDM) RUM (Reparametrized Unified Model) to compare cognitive structure of Q matrices elaborated in various conditions by a panel of experts, for two mathematics' multiple choice tests. In the first condition, only the items are available; in the second one, the items are presented as set together with a factor analysis; and in the third condition, the item difficulty and discrimination parameters are provided to experts as well as the distractors' analysis results. The ways that RUM parameters are related to the cognitive structure and are estimated, the difficulties met and the choices made, as well as their

## ABSTRACTS OF THE CONTRIBUTIONS

---

implications, are presented. In addition, some comparisons of the Q-matrices' cognitive structures between conditions, based on estimated RUM parameters, are presented and discussed.

**Lu, Z.**, and Yuan, K.-H.  
University of Notre Dame, South Bend, IN (zlu@nd.edu)

Poster-Tuesday-pm-s50

### Robust Procedures for SEM with Missing Data

Data for SEM analysis are typically collected through questionnaire and often contain outliers and missing values. Both create problems for estimation and statistical inference. The multivariate normal distribution based ML has been developed for SEM with missing data. However, a single outlier can rend the result meaningless. Robust procedures have been developed for SEM with complete data, but they cannot be applied to practical data with missing values. This paper obtains robust SEM estimates through multivariate t based ML with missing data. It (1) develops algorithms for parameter estimation, (2) develops procedures for statistical inference, in particular model evaluation, (3) develops program code so that applied researchers can use the methodology. We study three methods: Direct ML, Two-stage ML, and Two-stage GLS. Standard error of parameters will be obtained using the so-called sandwich type covariance matrix or bootstrap method. Overall model evaluation will be using likelihood ratio, rescaled likelihood ratio, and etc. A CFA model is illustrated through a real dataset (Mardia et al, 1979). Their robust estimates and model inference (consistent SE and test statistics) will be obtained and compared under various missing mechanisms: MCAR, MAR and NMAR.

**Magidson, J.**, and Vermunt, J.K.  
Statistical Innovations Inc., Belmont, MA (jay@statisticalinnovations.com)

IS-Tuesday-am-s25

### Imposing a Factor Structure on the Latent States in Hidden Markov Models to Improve Interpretability and Model Fit

A common strategy for determining the number of latent states in hidden Markov models (HMM) is to estimate successive models starting with a Markov process with  $K=2$  latent states, followed by 3 states, and continue to increment  $K$  by 1 until an acceptable fit to the data is found. An alternative approach, explored in this paper, is to begin with a 2-state process, but then each successive model adds an additional process (an additional dimension) rather than an additional state. For example, two uncorrelated 2-state processes with states denoted by  $S1 = (1,2)$  and  $S2 = (1,2)$  yields a total of 4 states  $S = \{(1,1), (1,2), (2,1), (2,2)\}$ . This approach results in parsimonious models that often provide improved fits to data. We illustrate the approach using longitudinal data on the severity of schizophrenia measured over 7 time points. The results suggest that the transition probability from a more severe to a less severe state decreases for those patients receiving an experimental drug relative to those in the control group. Interestingly, the significant treatment effect exists primarily on just one of the processes, leading to a more clearly defined definition of the treatment effect.

Frederickx, S., **Magis, D.**, Tuerlinckx, F., and De Boeck, P.  
K.U. Leuven, Belgium (david.magis@psy.kuleuven.be)

TS-Tuesday-am-s38

### A Crossed Random Effects Model to Detect Differential Item Functioning

We present a new methodology for detecting differential item functioning (DIF) in tests. We introduce a mixture DIF model which provides a clear-cut classification of items in two latent classes, namely a DIF and a non-DIF class. We consider both the item and person parameters to be random, thus dealing with a crossed random effects model. The main advantage of this model is that it does not require a set of anchor items to identify DIF items. We evaluate the performance of the mixture DIF model in simulation studies and compare this with traditional procedures. Extensions to more general settings are also discussed.

**Malone, P.S.**, Masyn, K.E., Lamis, D.A., and Northrup, T.  
University of South Carolina, Columbia, SC (malone.ps@gmail.com)

CS-Wednesday-am-s63

### Parallel-Process Discrete-Time Survival Modeling Via Latent Transition Analysis

Discrete-time survival modeling for a single outcome process (e.g., school dropout, recidivism, relapse, death) is well-understood. However, the application of survival modeling for two concurrent time-to-event processes has not yet been established for discrete-time models. Unlike a traditional competing risk model where the occurrence of one type of event precludes the occurrence of other type of events, in the proposed model occurrence of one event does not terminate risk for the other events. Further, the onset of each event may serve as a predictor of the subsequent onset of the other event. The model is formulated as a cross-lagged discrete-time Markov Chain. This particular parameterization facilitates the investigation of the temporal sequencing of the event processes that have already been shown to be correlated. It also allows the modeling of a mediation process where the influence of a given risk factor on one event process is mediated by the other event process. The presentation of this modeling approach includes an illustration with data on a worked example (school dropout and initiation of drug use) estimated using known-class mixture models in the *Mplus* v.5 software to construct a latent transition model estimating the effects of each event on the hazard of the other. Preliminary simulation results for the method will also be presented. Discussion includes extension to models of one-time events as mediators and competing-hazard models, in which the onset of one event (e.g., death) precludes the other.

**Mao, M.-M.**, Ding, S., Chen, Q., and Zhu, Y.-F.  
Jiang Xi Normal University, Nan Chang, China  
(ding06026@163.com or jie\_fang1@yahoo.com.cn)

TS-Tuesday-pm-s48

### New Classification Methods in Attribution Hierarchy Model

Leighton et al. (2004) proposed a cognitive diagnostic model named Attribute Hierarchy Method (AHM). There are two kinds of classification methods with AHM: method A and method B. Some defects of the two methods are analyzed. Several kinds of new classification methods based on AHM are proposed. The new classification methods proposed are based on establishing a series of indices of the similarity between the expected response pattern (ERP) and the observed response pattern (ORP). In order to promote the precision of the classification, two probabilities are considered carefully under the local independence: the first is  $L1$ , i.e., the likelihood of ERP (or ORP); and the second is  $L2$ , i.e., the probability of coinciding part of ERP with ORP. The similarity index is defined as  $L1/L2$ . Then a series of new classification methods are developed by combining the similarity index and the two kinds of the probabilities mentioned above. The performance of all the methods mentioned above has been assessed through the simulation research under the condition used by Ying Cui et al. (2006) The results of simulation indicate that the new methods are better than methods A and B. And with the slips increasing the advantage is especially obvious.

**Maris, G.**<sup>(1)</sup>, and van der Maas, H.L.J.<sup>(2)</sup>  
<sup>(1)</sup> Cito, Arnhem and University of Amsterdam, The Netherlands  
(gunter.maris@cito.nl or G.K.J.Maris@uva.nl)  
<sup>(2)</sup> University of Amsterdam, Amsterdam, The Netherlands

IS-Wednesday-am-s61

### Scoring Rules Based on Response Time and Quality

In 2005 van der Maas and Wagenmakers published an article in the American Journal of Psychology on the Amsterdam Chess Test. The Amsterdam Chess Test measures chess playing proficiency through a number of tasks. The Amsterdam Chess Test is scored using a scoring rule, called the correct item summed residual time, that depends both on response time and response quality. The items in the Amsterdam Chess Test have a deadline for responding and the score is the sum of the residual times for all items that are answered correctly. That is, incorrect responses are neither rewarded nor punished, and for correct responses the reward increases if their response time decreases. In this

## ABSTRACTS OF THE CONTRIBUTIONS

---

presentation we argue that this scoring rule is problematic, because for less able chess players guessing is an effective strategy to increase their (expected) score. For this reason a more general scoring rule, incorporating punishment for incorrect responses, is introduced. Furthermore, the exponential family IRT model for which this score is the sufficient statistic for ability is derived and it is shown how this new IRT model relates to the two-parameter logistic IRT model.

**Masyn, K.E.**

IS-Tuesday-am-s34

University of California, Davis, CA (kmasyn@ucdavis.edu)

### Modeling Multiple Mechanisms for Zeroes in Longitudinal Processes

Often in longitudinal studies of non-normative or aberrant behaviors, researchers are confronted with data that are highly skewed by absence of the behavior in a significant portion of the sample at any given time point. *Two-part* modeling is one approach that has been taken which allows modeling of a zero-part (whether or not an individual engages in the behavior or not at a certain point in time) simultaneously with a growth-part (modeling the level of behavior at a certain point in time, presuming the level is non-zero). However, these models do not allow for the possibility of multiple mechanisms that generate the observed zeroes. At any given time point there may be zero outcomes for a subset of subjects that have never engaged in the behavior and for a different subset that have previously engaged in the behavior but have desisted. The mechanisms that predict initiation may be distinct from the mechanisms that predict desistence following onset. This presentation discusses methodology for modeling multiple mechanisms for zeroes in longitudinal processes in a comprehensive latent variable framework. An approach that combines survival analysis (predicting time of onset), and two-part growth modeling (describing behavior trends and desistence patterns following initiation), is illustrated.

**Mavridis, D.**

IS-Monday-am-s6

University of Edinburgh, Edinburgh, United Kingdom (d.mavridis@ed.ac.uk)

### Goodness-of-Fit Tests and Detection of Atypical Response Patterns in Latent Variable Models Using Posterior Predictive Checks

Goodness-of-fit methods and detection of outliers in latent variable models have not received much attention. For categorical data, overall and limited information goodness-of-fit statistics together with their asymptotic distributions have been developed while person fit statistics have been suggested for detecting aberrant response patterns. In this paper, we explore Bayesian counterparts of those statistics by positioning the observed value of a statistic on its posterior distribution. This is achieved by posterior predictive checks which condition on the hypothesized model and compare the value of a statistic in terms of the values it would have assumed had the study been replicated using the same model. Replicated data sets, generated under the hypothesized model, should resemble the observed data had the model been correct. This method is straightforward when MCMC is used for estimating model parameters and is very useful for complex models for which there are no universally accepted statistics for evaluating model fit.

Joe, H.<sup>(1)</sup>, and **Maydeu-Olivares, A.**<sup>(2)</sup>

CS-Wednesday-am-s62

<sup>(1)</sup> University of British Columbia, Vancouver, Canada

<sup>(2)</sup> University of Barcelona, Barcelona, Spain (amaydeu@ub.edu)

### Constructing Chi-Square Goodness-of-Fit Tests for Multinomial Data that Are more Powerful than Pearson's $\chi^2$

Maydeu-Olivares and Joe (2005, 2006) introduced classes of chi-square tests for (sparse) multidimensional multinomial data based on low-order marginal proportions. Their statistics were shown empirically in some cases to have better power than Pearson's  $\chi^2$  statistic. We extend their work by providing general conditions under which

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

quadratic forms in summary statistics of cell proportions are asymptotically chi-square (even for one-dimensional multinomials). A main contribution is to show theoretically why the class of statistics has good power, even for non-sparse tables. With these results, we show how quadratic-form statistics can be constructed that are more powerful than  $X^2$  and yet, have null distribution can be well approximated by a chi-square distribution in finite samples with large models. An example with binary item response data is used to illustrate the theory.

**Mayekawa, S.I.**

CS-Tuesday-am-s27

Tokyo Institute of Technology, Tokyo, Japan (mayekawa@nifty.com)

### Estimation of Ability Using the Globally Optimal Scoring Weights

In IRT, the best scoring weights for calculating the ability parameter ( $\theta$ ) is known to be dependent on  $\theta$ . In this research, for the polytomous models, several scoring weights were derived by maximizing the expectation of the information function with respect to the reference  $\theta$  distribution. The ability parameters were calculated as the Bayesian posterior mean given the weighted scores, and the resulting posterior standard deviations were compared to the usual procedures. It was found that, if we choose a correct reference  $\theta$  distribution, the results were comparable.

**McIntyre, H.H.**, and Embretson, S.E.

Poster-Tuesday-pm-s50

Georgia Institute of Technology, Atlanta, GA (hhayes3@gatech.edu)

### Application of Profile Analysis Using Multidimensional Scaling (PAMS) to a Battery of Emotional Intelligence Tests

Profile analysis has been widely used in education and clinical psychology for classifying individuals in terms of strengths and weaknesses across an array of cognitive ability domains and the relative potency of particular psychological disorders, respectively. Although several techniques are available for carrying out a profile analysis, Profile Analysis Using Multidimensional Scaling (PAMS) has gained momentum in recent years due to its computationally efficient method of identifying major (i.e., dominant) profiles within a population as well as determining an individual's degree of similarity to each major profile (e.g., Kim, Davison, & Frisby, 2007). The purpose of the current analysis was to apply PAMS to a sample of  $n = 200$  participant scores on a battery of Emotional Intelligence tests that included both objectively-scored and self-report measures. Results from the Multidimensional Scaling (i.e., exploratory) portion of the analysis support the presence of two major profiles, and a subsequent Structural Equation Modeling analysis confirmed these findings. Of particular interest is the first major profile, which reflects a tendency for individuals to report themselves as being more emotionally intelligent than their objective scores indicate. Suggestions for testing the criterion-related validity of profile similarity scores are discussed.

**Merkle, E.C.**

CS-Tuesday-am-s35

Wichita State University, Wichita, KS (edgar.merkle@wichita.edu)

### Binary Recursive Partitioning: Uses in Psychology

Binary recursive partitioning (BRP), also known as classification and regression trees (Breiman et al., 1984), is a computationally intensive, nonparametric statistical method that can be a useful alternative to regression/ANOVA. Instead of imposing a stochastic model on the data, BRP identifies binary splits of predictor variables that are useful for predicting the response variable. No significance tests are involved. While there have been a small number of BRP applications/developments in psychology, the method is not well-known in the field and could be useful to many researchers. In the presentation, I describe BRP methods and their application to psychology. I also discuss BRP's relevance to the ongoing discussion regarding problems with significance testing.

**Millsap, R.E.**, and Lee, S.  
Arizona State University, Tempe, AZ (millsap@asu.edu)

CS-Monday-am-s9

### Approximate Fit in SEM Without A Priori Cutpoints

Recent research has questioned the generality of some of the cutpoints in standard use for approximate fit indices in SEM. A general problem is that over different model structures and data conditions, it is difficult to determine what types of model misspecifications are consistent with a given level of approximate fit. As a consequence, it is difficult to specify cutpoints for common indices of fit. Here we present a simulation-based method for determining whether the fit of a given model is consistent with particular forms of misspecification. The method does not require the investigator to specify a priori cutpoints for the approximate fit index of interest. The method is described and an example of an application is given.

**Miyazaki, K.**, Hoshino, T., and Shigemasu, K.  
The University of Tokyo, Tokyo, Japan (miyazaki@bayes.c.u-tokyo.ac.jp)

CS-Tuesday-am-s29

### A Bayesian Semiparametric Item Response Model with Dirichlet Process Priors

In the Item Response Theory (IRT), item characteristic curves are expressed with logistic models or normal ogive models. However, because only limited patterns of shapes can be obtained from logistic models or normal ogive models, there is a possibility that the model assumed doesn't fit the data and as a result, the existing method can be rejected because it cannot deal with various item response patterns. A lot of nonparametric IRT models without distributional assumption have been proposed to solve this problem, but since nonparametric IRT models don't include the ability parameters explicitly, they cannot calculate scores of examinees. To overcome the problems described above, we proposed a new semiparametric IRT model using Dirichlet process mixture logistic distribution. Our method needs no assumption except for that item characteristic curves are monotonically nondecreasing function and can estimate the ability parameters directly, which nonparametric IRT methods cannot. The results of two simulation studies indicate that the proposed method can express more patterns of shapes for item characteristic curves and can estimate the ability parameters more accurately than the existing parametric or nonparametric method. The proposed method was also applied to the Facial Expression Recognition data and meaningful results were obtained.

**Miyazaki, Y.**  
Virginia Polytechnic Institute and State University, Blacksburg, Virginia  
(yasuom@vt.edu)

Poster-Tuesday-pm-s50

### Latent Effects Factor Model

A model that factors regression slopes at level 1 in two level hierarchical linear models will be presented. Multilevel factor analysis works for factor analyzing the means, i.e., random intercepts, but the proposed model, termed as latent effects factor model, works for the random slopes. In behavioral and social sciences, the ultimate interests of substantive researchers lie at regression slopes that indicate the impact of the independent variables on the dependent variable or the strength of association between those. When we have multiple independent variables in many macro units (i.e., groups) such as schools and organizations, these sets of regression slopes which possibly vary from group to group can be regarded as indexes that characterize the unique nature of each group. There are many cases that these regression slopes covary each other with relatively high correlations. In such cases, it is possible to imagine factors operating at the macro levels that generated the shared variation. The factorization of regression slopes may facilitates a better interpretation of the results by reducing the number of parameters in the model. The worked-out examples from sociology of education will be used to illustrate the techniques and possible interpretations of the results.

**Molenaar, D.**, and Dolan, C.V.  
 University of Amsterdam, Amsterdam, The Netherlands  
 (D.Molenaar@uva.nl)

CS-Wednesday-am-s55

Factor Analytic Modeling of Ability Differentiation

In intelligence research, a well established finding is that scores on various cognitive ability tests are strictly positive correlated, even though they concern distinct cognitive abilities. This is referred to as the positive manifold. An explanation for the positive manifold is that a general intelligence factor  $g$  underlies all cognitive tests. It is often hypothesized that the positive manifold is stronger within subjects low in  $g$  and weaker within subjects high in  $g$ . This phenomenon is known as ability differentiation. In past research, ability differentiation was demonstrated using correlations, principal components and factor analyses. However, no attempt was made to find a specific locus of the differentiation effect in the common factor model. In the present talk, we present factor analytic models that are suitable to test specific hypothesis about the locus of the hypothesized differentiation effect.

**Mooijaart, A.**, and Satorra, A.  
 University of Leiden, Leiden, The Netherlands (mooijaart@fsw.leidenuniv.nl)

CS-Monday-am-s9

Does the Normal Theory Test Statistics Always Detect Nonlinear Terms in Structural Equation Modeling?

Consider in the population that the following regression equation holds

$$y = \bar{\beta}_0 + \bar{\beta}_1 \xi_1 + \bar{\beta}_2 \xi_2 + \bar{\beta}_{12} \xi_1 \xi_2 + \zeta, \quad (1a)$$

where there are several indicators for the latent variables ( $\xi_1$  and  $\xi_2$ ). In this formulation no assumption is made about the statistical distribution of the latent variables and the disturbance ( $\zeta$ ). The only assumption we make is that the disturbance is independent of the factors, i.e., not just uncorrelated. Furthermore, each factor has two indicators. So besides the regression equation as given in (1a), the following measurement equations are specified

$$\begin{aligned} x_j &= \bar{\alpha}_j + \bar{\lambda}_j \xi_1 + \delta_j \text{ for } j = 1, 2 \\ x_j &= \bar{\alpha}_j + \bar{\lambda}_j \xi_2 + \delta_j \text{ for } j = 3, 4. \end{aligned} \quad (1b)$$

The null model,  $M_0$ , is specified as

$$M_0 : y = \beta_0 + \beta_1 \xi_1^* + \beta_2 \xi_2^* + \zeta^*. \quad (2)$$

In model  $M_0$  it is assumed that the vector  $(\xi_1^* \quad \xi_2^* \quad \zeta^*)'$  is trivariate normal distributed with zero correlations between the disturbance and the factors. The question is: Do we get a hint that  $M_0$  is misspecified if we analyze data under  $M_0$ , where the data come from the population for which (1a) and (1b) hold. In other words: does the NT test statistic has the capacity to detect the interaction term? This is of practical importance, because if there is not such an indication a researcher will incorrectly conclude that there is just a linear relationship between the variables.

Rizopoulos, D.<sup>(1)</sup>, and **Moustaki, I.**<sup>(2)</sup>  
<sup>(1)</sup>K.U. Leuven, Belgium

IS-Monday-am-s6

<sup>(2)</sup> London School of Economics, London, United Kingdom (i.moustaki@lse.ac.uk)

### Latent Variable Models with Non-Linear Effects

Latent variable models with higher-order and interaction terms have recently been developed in the literature. Most of that work is within the area of structural equation modeling. In this paper, we consider a generalized latent variable model framework for mixed responses (metric and categorical) that allows the inclusion of both nonlinear latent terms and covariate effects. The model parameters are estimated using full Maximum Likelihood based on a hybrid integration-maximization algorithm. An application and a simulation study are used to illustrate the methodology proposed.

**Murakami, T.**

IS-Tuesday-am-s26

Chukyo University, Toyota, Japan (tandem06@sass.chukyo-u.ac.jp)

### Individual Differences in Three-Way Rating Scale Data and their Description by Three-Mode PCA

The distinction between three-way rating scale data and three-way profile data (Kroonenberg, 2008) is very important in the application of three-mode principal component analysis (TMPCA) because they require different ways of preprocessing of raw data and of interpretations of results. A typical example of three-way rating scale data set is obtained by the semantic differential method, and we will concentrate on this kind of data here. The essence of the analysis is the separate decomposition of an individual two-way data matrix consisting of ratings of objects on scales by each participant into the product of a score matrix for objects and a loading matrix for scales, and TMPCA just represents variations in individual score and loading matrices in more parsimonious form. Several aspects of individual differences among participants are distinguished, e.g., locations of objects on cognitive and affective dimensions, amounts of variations attributed to response sets, and proportions of explained variances. TMPCA shows some aspects in the output directly while additional indices are needed for others. An illustrative example using a real data set with newly devised graphical representations will be presented.

**Murakami, T.**

CS-Tuesday-am-s37

Chukyo University, Toyota, Japan (tandem06@sass.chukyo-u.ac.jp)

### Prototype Transformations of Principal Components for Developing Psychometric Scales

Prototype transformation is a simple (extended) oblique rotation procedure of principal components (Murakami, 2007). It defines the transformed component as a unit-length vector passing through the prototype item, which is specified by the researcher in the substantive domain as the most typical one to measure the intended constructs. Two small devices for using the procedure to develop or improve psychometric scales defined as simple sums of item responses will be proposed; an equation to obtain a structure matrix of transformed components from unrotated component loadings, and a simplest rule to classify each item into the subscale including the prototype item, the component of which is closest to it. While prototype transformation may be expected to be a tool to introduce incomplete theoretical ideas into essentially exploratory data analysis, some difficulties in applications will be explained through numerical examples using real data.

**Asparouhov, T., and Muthén, B.**

KN-Monday-pm-S23

University of California, Los Angeles (bmuthen@ucla.edu)

### Exploratory Structural Equation Modeling

Exploratory factor analysis (EFA) has been said to be the most frequently used multivariate analysis technique in statistics. In 1966 Jennrich solved a significant EFA factor loading matrix rotation problem by deriving the direct quartimin rotation. He was also the first to develop standard errors for rotated solutions although these have still not made their way into most statistical software programs. This is perhaps because Jennrich's achievements were partly overshadowed by the 1967 development of confirmatory factor analysis (CFA) by Joreskog. Joreskog developed CFA further into structural equation modeling (SEM) where CFA was used for the measurement part of the model. The strict requirement of zero cross-loadings in CFA, however, often does not fit the data well and has led to a tendency to rely on extensive model modification to find a well-fitting model. In such cases, searching for a well-fitting measurement model may be better carried out by EFA (Browne, 2001). Furthermore, misspecification of zero loadings tends to give distorted factors with over-estimated factor correlations and subsequent distorted structural relations. This paper describes an EFA-SEM (ESEM) approach, where in addition to or instead of a CFA measurement model, an EFA measurement model with rotations can be used in a structural equation model. The ESEM approach has recently been implemented in the Mplus program. ESEM gives access to all the usual SEM parameter and the loading rotation gives a transformation of structural coefficients as well. Standard errors and overall tests of model fit are derived. Geomin and target rotations are discussed. Examples of ESEM models include multiple-group EFA with measurement and structural invariance testing, test-retest (longitudinal) EFA, EFA with covariates and direct effects, and EFA with correlated residuals including EFA of traits in MTMM settings. Testing strategies with sequences of EFA and CFA models are discussed. Simulated and real data are used to illustrate the points.

**Nakamura, K.**

CS-Monday-pm-s20

Saitama Gakuen University, Saitama, Japan (kiike@toki.waseda.jp)

### An Analysis of Students' Evaluations of Teaching via a Multigroup Latent Mixture Graded Response Model

Students' evaluations of university teaching play an important role to lead to the improvement of teaching. In this study, a multigroup item response model is applied to rated data with a five-point scale in order to compare the results of ratings for each teacher on a common scale. In addition, we detect differences of students' rating styles by assuming latent classes.

**Nishisato, S.**<sup>(1)</sup>, and Clavel, J.G.<sup>(2)</sup>

CS-Wednesday-am-s56

<sup>(1)</sup> University of Toronto, Canada (snishisato@oise.utoronto.ca)

<sup>(2)</sup> Universidad de Murcia, Spain

### Total Information Analysis: An Extension of Dual Scaling

Nishisato and Clavel (2008) proposed total information analysis by extending dual scaling, correspondence analysis and homogeneity analysis to both within-set and between-set distances and by looking at data structure in total space, hence the name total information analysis (TIA). The first paper offers elaboration of this extension, with logical and numerical examples, to show how reasonable and sound this approach is. In comparison, the traditional quantification procedure is based solely on within-set analysis, thus ignoring between-set relations, and covers this deficit by the so-called symmetric scaling. It is our urgent appeal to the current research on quantification theory that TIA should replace the traditional approach since the latter lacks logical rigor. Why this is so will be thoroughly explained.

**Noel, Y.**

IS-Monday-pm-s17

Université de Rennes 2, Rennes, France (yvonnick.noel@uhb.fr)

### The Beta Unfolding Model for Continuous Bounded Responses

This paper aims at providing an unfolding model for continuous bounded responses, by which are designated continuous responses with both a lower and an upper bound. Such data may be collected by asking subjects to rate items by putting a mark at some point on a horizontal line segment, which ends are labeled “0% agreement” and “100% agreement”, for instance. The response is measured as the distance from the left end of the segment to the subject's mark. Assuming an interpolation response mechanism (Noel & Dauvier, 2007), a beta distribution is theoretically deduced as a model for such responses. The two natural parameters of the beta are interpreted as acceptance and refusal parameters, and expressed as functions of person-item distances on some latent continuum. The bimodality feature of the beta is used to model chaotic response choice among ambivalent subjects, and a parallel with “cusp” models is discussed. An EM estimation algorithm is described that performs well in a simulation study. The model is applied to attitude toward abortion data.

**Oh, H.**

TS-Tuesday-am-s31

Educational Testing Service, Princeton, NJ (hoh@ets.org )

### Multi-Group Confirmatory Factor Analysis: To Evaluate Construct Comparability

There are four conditions that must be met for test equating: the test should measure the same construct as well as exhibit equity, symmetry, and there should be population invariance. When a test undergoes substantial changes, it is necessary to establish that the changes have not had a significant impact on the constructs measured by the test. Thus, in order to qualify as an equated test, it is essential that the previous version of a test and a newly configured test should measure the same constructs. It is often acknowledged that the data collection design and statistical procedures used to equate different forms of the same test can be used to scale scores from tests with different attributes. However, such procedures do not guarantee that those scores are comparable or interchangeable. The purpose of this study is to provide some indication of the degree of factorial similarities and differences between the constructs measured by newly configured large scale test and constructs measured by the old version of the test across gender subgroups using confirmatory factor analysis. The results suggest that the changes to the test have had very little impact on the dimensional structure of the test.

**Okada, K.**, and Shigemasu, K.

CS-Monday-am-s7

The University of Tokyo, Tokyo, Japan (ken@bayes.c.u-tokyo.ac.jp)

### Confirmatory Multidimensional Scaling: A Model Selection Using Bayes Factors

Inequality constraints among objects can be used to assign a specific meaning to the multidimensional scaling model. Researchers often have one or more theories or expectations with respect to the outcome of their empirical research. Different models arise if different sets of constraints are used. We have previously proposed a method for estimating distance-constrained multidimensional scaling model using Markov chain Monte Carlo method. In this paper, model selection using Bayes factors is discussed. We show that traditional maximum likelihood based information criteria sometimes fail in evaluating models with inequality constraints. Therefore we propose the use of Bayes factors for model evaluation. We illustrate the proposed method using artificial data. Then, the example of confirmatory multidimensional scaling is presented using empirical data.

**Okubo, T.**<sup>(1)</sup>, Hoshino, T.<sup>(2)</sup>, and Mayekawa, S.I.<sup>(3)</sup>

CS-Monday-pm-s20

<sup>(1)</sup> The National Center for University Entrance Examinations  
(NCUEE), Tokyo, Japan (okubo@rd.dnc.ac.jp)

<sup>(2)</sup> Nagoya University

<sup>(3)</sup> Tokyo Institute of Technology, Tokyo, Japan

### Order-Constrained Nominal Categories Model

Several models have been presented to analyze the ordered response data in item response theory. One of them is a group which is classified as the divide-by-total model which includes Master's (1982) partial credit model (PCM) and Muraki's (1992) generalized partial credit model (GPCM). Those models can be characterised as the model which defines the ICRF of the category proportional to the exponential of the linear function of the ability, and can be thought of the sub models of Bock's (1972) nominal categories model (NCM). In GPCM, the differences between the slope parameters of the neighboring categories are constant, whereas in the PCM, all slope parameters are constrained to one. In this research, following Samejima's (1969) suggestion, we propose the order-constrained nominal categories model (ONCM), which is a sub model of NCM where the slope parameters are ordered. Since the constraint of slope parameters for ONCM is more relaxed than the ones for PCM or GPCM, we can expect a better model fit. The marginal maximum likelihood estimation of ONCM was developed and the performance of ONCM was compared to the other ordinal response model analyses.

**Olsson, U.H.**, Foss, T., and Jöreskog, K.G.

CS-Monday-am-s9

Norwegian School of Management, Sandvika, Norway (ulf.h.olsson@bi.no)

### The Power of the Non-Normality Corrected Chi-Square Statistics in Structural Equation Modeling

In this paper we examine the power of two “chi-square” statistics, namely the SB statistics (Satorra & Bentler, 1988, 2003) and the ADF statistics (Browne, 1984). The SB statistic corrects the normal theory chi-square with a scale factor which is estimated from the sample and involves the estimated asymptotic covariance matrix (ACM). The scale factor is estimated so that the SB statistic has an asymptotically correct mean. The ADF statistic under the assumption of correct model has an asymptotic chi-square distribution. We will demonstrate how the power of SB and ADF varies with increasing kurtosis in a homogeneous – and a non-homogeneous way. Since both SB and ADF depend on kurtosis the respective chi-square statistics are affected by kurtosis. We show that these “chi-square” statistics tend to decrease with increasing kurtosis. The practical consequence of this is that models that do not hold tend to be accepted by these “chi-square” tests if kurtosis is large. Although the results developed here can be demonstrated by simulations and analyzing random samples, we will use a different approach. Simulation studies depend on rather arbitrary conditions of the design of the simulation. By contrast our results are valid under fairly general conditions.

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

**Otsu, T.**, Ishioka, T., and Hashimoto, T.

Poster-Tuesday-pm-s50

The National Center for University Entrance Examinations (NCUEE),  
Tokyo, Japan (otsu@rd.dnc.ac.jp)

### Building a Statistical Database of NCT Test Items

The National Center Test (NCT) is a nationwide university entrance examination organized annually by the National Center for University Entrance Examinations (NCUEE) and Japanese universities. All national and public universities in Japan have been adopting the NCT. In addition, many private universities use the NCT as one of the tests to screen students. Last January, about 500,000 participants took the NCT. Although the item development procedure of the NCT is a highly comprehensive process involving reviews by anonymous professors, the NCUEE does not use modern test theory to moderate difficulties. Organizing statistical data of the NCT is key to item writing for future examinations and educational policy making. We developed a database of item statistics of the NCT between 1990 and 2007. This database contains several statistics based on classical test theory. We included about 18,000 items from various areas. The data is contained in XHTML forms, which are generated using Prolog based XML parser, with links to scripts for dynamic chart generation. The forms also contain links to item documents for reference. The statistics reveal difficulties in moderating scores in the absence of pre-test operations.

**Paccagnella, O.**

TS-Tuesday-am-s38

University of Padua, Padua, Italy (omar.paccagnella@unipd.it)

### Anchoring Vignettes with Sample Selection

Anchoring vignettes (King et al., 2004) are used to analyse ordinal survey responses taking into account individual differences in interpretation of the survey questions. Vignettes are a new tool for enhancing self-report data comparability across countries. Many results support the ability of the vignettes to correct for differential item functioning (DIF). SHARE (Survey of Health, Ageing and Retirement in Europe) is a new survey which collects data on individual life circumstances of 50+ people in 14 European countries, using a CAPI program and supplemented by a self-completion paper and pencil questionnaire drop-off. In some countries, a sample of individuals was randomly selected to receive vignettes as drop-off. Unlike other surveys working with vignettes, in SHARE a sample selection problem may arise also when respondent completes the CAPI questionnaire, but refuses to fill in the vignette drop-off. Fitting models to the observed sample ignoring potential selection bias may lead to inconsistent estimates. This paper aims at extending the standard model for estimating vignettes in order to allow the specification of some selection variables. In our model, a wide set of explanatory and selection variables are available (demographic, socio-economic, cognitive, physical and mental health information). Results are then compared across countries.

**Papanastasiou, E.C.**, and Reckase, M.D.

IS-Monday-pm-s18

University of Nicosia and European Psychometric Services, Nicosia,  
Cyprus (papanastasiou.e@unic.ac.cy)

### Item Review as a Non-Traditional Method of Item Analysis

If good measurement depends in part on the estimation of accurate item characteristics, it is essential that test developers become aware of discrepancies that may exist on the item characteristics before and after item review. It is also essential for them to determine whether it is appropriate to use the item characteristics of items from testing situations that allowed item review, to situations that do not allow item review, or vice versa. The results of this study will enable us to obtain answers to such issues. The research questions that will be examined in this study include; How do the item parameters of the 3PL model change before and after item review? Can examining patterns of item changing provide indications on the poor quality of some test items? Finally, what is the effect of item review on the examinee's ability estimates when one takes into account the changes in the item parameters due to item review?

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

**Parker, P.,** Keller, R., and Keller, R.  
University of Massachusetts, Amherst, MA  
(paparker@educ.umass.edu)

TS-Monday-pm-s22

### The Examination of the Classification of Students into Performance Categories by Two Different Equating Methods

The purpose of the following study is to provide a comparison of two item response theory equating methods: True Score Equating and Estimated True Score Equating (van der Linden, W. J., 2000). IRT true score equating is commonly used in many testing situations whereas the local equating is a recently proposed method that has not been widely used. The study examines the differences in the classification of students into performance categories using each of the methods. As both equating methods rely upon IRT, they require that the item parameters first be placed onto a common scale. This can be accomplished with a variety of scaling methods. Therefore, the effect of the scaling method chosen will also be investigated in this study, by examining any differences in the classifications of examinees due to different implementations of the fixed common item parameter (FCIP) scaling method, namely, FCIP-1 and FCIP-2. Implications for operational testing companies are discussed.

**Bartolucci, F.,** and **Pennoni, F.**  
University of Milano-Bicocca, Milano, Italy (fulvia.pennoni@unimib.it)

IS-Tuesday-am-s25

### The Latent Markov Rasch Model

We introduce a model for the analysis of binary data deriving from the administration of test items at one or more occasions. The model assumes that the behavior of each subject depends on a discrete latent process which, as in the latent Markov model, follows a first-order Markov chain. The proposed model also adopts a Rasch-type parameterization of the distribution of the response variables given this process. For the maximum likelihood estimation of the model parameters we illustrate an EM algorithm implemented by means of certain recursions known in the hidden Markov literature. The algorithm can also be applied with many time occasions. We also illustrate how hypotheses of interest on the transition matrix may be tested by using a likelihood ratio statistic between nested models. The approach is extended to the case of: (i) multivariate longitudinal data; (ii) individual covariates. In the first case we assume conditional independence between the response variables at the same occasion given the latent process. In the second case, we allow the individual covariates to affect the initial and the transition probabilities of the latent process. The approach is illustrated by an application involving educational data and another application related to a psychological experiment.

**Polak, M.,** Heiser, W.J., and de Rooij, M.  
Leiden University, Leiden, The Netherlands (polak@fsw.leidenuniv.nl)

IS-Monday-pm-s17

### A Comparison of Correspondence Analysis with Parametric and Nonparametric IRT Models for Analyzing Single-Peaked Responses

Single-peaked responses naturally arise in a variety of research settings, such as ecology (e.g., Ter Braak & Prentice, 2004) or attitude measurement (e.g., Roberts, Donoghue, & Laughlin, 2000). In the field of ecology the aim is to scale species and sites, which is often accomplished with Correspondence Analysis (CA). When single-peaked data are strongly one-dimensional, a two-dimensional CA representation will show the well-known “arch-effect”, where the sites (items) and species (persons) are ordered along an arch (but also on the first dimension) according to their position on the scale (e.g., Hill, 1974). In psychometrics, an increasingly popular approach to scaling persons and items is unfolding item response theory (unfolding IRT). In the current paper we compare CA with both a parametric and a nonparametric unfolding IRT model: GGUM (Roberts, Fang, Cui, & Wang, 2006) and MUDFOLD (Van Schuur & Post, 1998), respectively. In particular, we explore the surplus value of different types of data coding in correspondence analysis, namely conjoint and disjoint coding (which is multiple correspondence analysis), in relation to the item threshold-estimates in IRT unfolding. Furthermore, we show that the arch-effect is still present in

## ABSTRACTS OF THE CONTRIBUTIONS

---

CA with conjoint coding, but that CA with disjoint coding results in an arch with inwardly bending extremes, causing underestimation of the extreme person locations.

**Powell, J.C.**

CS-Tuesday-am-s27

Robert Morris University, Moon Township, PA (jpowell@tir.com)

### Observing Learning Using All Answers: A Commentary on Test-Scoring Practices

There is considerable pressure upon testing programs to provide more *formative* information than can be obtained from that provided by the total-correct scores on tests that is the current practice. This paper provides an alternative to this practice that uses *all the answers* that students provide to a multiple-choice test instead of only those that are classified as *right*. The system classifies all the students answers based upon their *reported reasoning* behind their answers selection and scales this reasoning in accordance with the *maturity level* displayed. The nature of the reasoning provides *formative* information, and the maturity level provides *summative* information so that both diagnosis learning characteristics and statements of learning progress can be obtained by the single scoring system. Application of the system to illustrative live data complements the presentation. The advantages and limitations of this system for establishing scoring protocols are presented.

**Price, L.**

Poster-Tuesday-pm-s50

Texas State University, San Marcos, TX (lrprice@mindspring.com or lrprice@txstate.edu)

### Deriving Optimal Neuroimaging Models Using Bayesian Model Averaging

The aim of this paper is to present a method for graphical model selection and optimization of the functional connectivity among regions of interest in the human brain where little or no previous theory exists. We propose a fully Bayesian approach for deriving network models among regions of interest that allows researchers to address the issue of uncertainty that is inherent in a “single-best model” strategy. We use the Occam’s Window algorithm of Madigan and Raftery, (1994) to identify the most plausible models, and then apply Bayesian model averaging (Madigan, Raftery, Volinsky, & Hoeting, 1996) to illustrate our method using data acquired by Positron Emission Tomography (PET) on a group of normal controls and a group exhibiting linguistic dysfunction under varying conditions of verbal activity. Our approach enables scientists to identify optimal models to examine different patterns of interregional causal effects due to an intervening condition within the network using Bayesian Structural Equation Modeling (BSEM).

**Rabe-Hesketh, S.**<sup>(1)</sup>, and Skrondal, A.<sup>(2)</sup>

IL-Monday-pm-s12

<sup>(1)</sup>University of California, Berkeley, CA (sophiarh@berkeley.edu)

<sup>(2)</sup>University of London, United Kingdom

### Comparison of Methods for Handling Endogenous Covariates in Longitudinal Data

A strength of longitudinal data is that they allow estimation of within-subject effects of time-varying covariates. Such estimates do not suffer from bias due to unobserved subject-specific covariates that are correlated with the observed time-varying covariates, a special kind of endogeneity. Within-subject effects are typically estimated by mean-centering the time-varying covariates, or similar approaches. An alternative is to specify a random intercept model as a structural equation model and allow the time-varying covariates to be correlated with the random intercept. In this talk, another alternative will be proposed. Some approaches will be shown to correspond to equivalent models in the sense that they are reparameterizations of each other. Furthermore, the equivalent models yield identical maximum likelihood estimates of within-subject effects. However, estimates of the effects of exogenous subject-specific covariates differ, and are consistent only for some of the approaches.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Raïche, G.**<sup>(1)</sup>, Magis, D.<sup>(2)</sup>, and Blais, J.-G.<sup>(3)</sup>

CS-Wednesday-am-s62

<sup>(1)</sup> Université du Québec à Montréal, Montréal, Canada

(Raiche.Gilles@uqam.ca)

<sup>(2)</sup> K.U. Leuven, Belgium

<sup>(3)</sup> Université de Montréal, Montréal, Canada

### Multidimensional Item Response Theory Models Integrating Additional Inattention, Pseudo-Guessing, and Discrimination Person Parameters

The result obtained by a person within a test is not always representative of its real level of proficiency and can constitute misleading information on its capacities. There are cases where the candidates miss attention, motivation or preparation and show underachievement. Their result does not correspond then any more to their true potential, an inappropriate response pattern being obtained. Indices of adjustment related with modelizations from item response theory were proposed to detect these candidates. However, those indices are unfortunately not integrated directly inside these modelizations and thus rest on an estimate of the level of proficiency which does not hold account of the inadequacy of the response pattern. To palliate to this situation new multidimensional modelizations of item response are proposed. Those add to the level of proficiency of the candidate, new person parameters of discrimination, pseudo-guessing and carelessness rather than of items only. These modelizations could make it possible to correct the value of the level of proficiency in the presence of factors which affect the adequacy of the person's response pattern.

**Rausch, J.R.**

IS-Monday-am-s5

University of Minnesota, Minneapolis, MN (rausch@umn.edu)

### Parametrizations for the Negative Exponential Growth Model

Researchers frequently depend on linear models for the analysis of longitudinal data. However, when attempting to precisely quantify and interpret change over time, nonlinear growth models can be more useful in a number of situations. The present talk focuses on a relatively simple nonlinear growth model, the negative exponential growth model. Various parameterizations of this model are available, providing a variety of options to researchers in the social and behavioral sciences. The primary purpose of this talk is to present new parameterizations of the negative exponential model, which often provide practical advantages over previous parameterizations of this model. An example is provided to illustrate the merits of the new parameterizations for models of negative exponential growth

**Raykov, T.**

CS-Monday-am-s10

Michigan State University, East Lansing, MI (raykov@msu.edu)

### Testing Multivariate Mean Collinearity With Missing Data: Potential for Enhanced Power in Group Mean Difference Analyses

A procedure for testing mean collinearity in multidimensional spaces is outlined that is applicable in settings with missing data. The approach is based on nonlinear parameter restrictions and is developed within the framework of mean and covariance structure modeling. The method provides useful information about multiple response centroid constellation, which can be especially helpful with designs characterized by data missing completely at random. The procedure is applicable when data are missing at random, can be associated with enhanced power when testing group mean differences simultaneously on several continuous outcomes, and can also be employed as a precursor to other multivariate multiple-population analyses. The proposed approach is illustrated with an example. Keywords: collinearity, mean and covariance structure modeling, missing data, multivariate analysis, nonlinear constraint.

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

**Reckase, M.D.**

IS-Monday-pm-s18

Michigan State University, East Lansing, MI (reckase@msu.edu)

If All Tests Measure Multiple Constructs, What Do they Measure Best?

Analysis of the typical tests of cognitive skills and knowledge show that they are complex instruments that need many different cognitive skills and types of knowledge to determine the correct responses to the test items. This implies that the tests differentiate among examinees in different ways depending on the particular set of skills and knowledge that the examinees have. This paper will consider how to describe what skills and knowledge a test is best at differentiating. Generally, this will be a weighted composite of skills and knowledge that depend on the skills and knowledge assessed by the individual items on the test. Both test and item characteristics will be described

**Reckase, M.D.**

CS-Tuesday-am-s36

Michigan State University, East Lansing, MI (reckase@msu.edu)

Addressing the Number of Dimensions Problem in Multidimensional Item Response Theory

Before multidimensional IRT (MIRT) analyses can be performed, it is necessary to decide on the number of dimensions to use for estimating parameters. This paper will make a distinction between the number of coordinate axes needed to model the relationships in the data matrix and the number of constructs being assessed by a test. The dimensionality problem will be defined in terms of the former rather than the latter. After providing a theoretical context for the problem, several of the commonly used methods for deciding on the number of dimensions for an exploratory analysis will be reviewed. These include DIMTEST/DETECT, parallel analysis, and the chi-square difference test for the difference in fit for  $n$  and  $n + 1$  dimensions. All of these methods were applied to simulated data sets designed to mimic real data and item response data from an operational testing program. The results show the accuracy of the recovery of the number of dimensions used to generate the simulated data and the variation of the results from the different methods for real data. A recommendation is developed for how to deal with this problem when applying MIRT procedures to the analysis of real data.

**Reininghaus, U., Kallert, T., Burns, T., Slade, M., Ruggeri, M., Hansson, L.,  
Croudace, T., and Priebe, S.**

IS-Monday-pm-s16

University of London, United Kingdom (u.reininghaus@qmul.ac.uk)

Improving the Measurement of Patient Reported Outcomes in Mental Health Research Using State of the Art Latent Variable Modeling Procedures

Over the past decade, patient reported outcomes (PROs) have become increasingly important in mental health research. However, there are a number of methodological challenges with regard to the psychometric properties and conceptual basis of PRO measures. Most importantly, evidence over the past decade has shown that there is a considerable overlap among widely used PROs such as subjective quality of life, needs for care, treatment satisfaction, or self-reported symptoms. The challenge therefore is to establish how the remaining construct-specific variance can be reliably captured independent from this overlap. The current research aims to assess the dimensional structure of existing PRO measures by applying exploratory and confirmatory factor analysis for categorical (ordinal) data in MPlus, Version 5, to a large pooled database of 6 European studies, which assessed PROs in people with severe mental disorder on the basis of widely used scales (e.g., Manchester Short Assessment of Quality of Life, Camberwell Assessment of Needs, Clients Scale for Assessment of Treatment). Several latent variable models are compared on the basis of their model fit indices, item discrimination parameters, and measurement precision. Preliminary findings indicate that the dimensional structure is best represented by a bi-factor model with one general and several construct-specific factors.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Ricks, R.A.**, and Sheng, Y.  
Southern Illinois University, Carbondale, IL (ysheng@siu.edu)

CS-Tuesday-am-s36

### An Investigation of Modified Stout's T Procedures for Small Datasets

Unidimensional IRT models rely on a strong assumption that each test item measures some facet of a unified latent trait. Violation of this assumption results in serious problems associated with the inference made about individual persons. In the literature, numerous indices assessing unidimensionality have been proposed. Among them, Stout's T, implemented in DIMTEST, was found to be more powerful than the other methods, but it does not work well in certain situations. Nandakumar and Stout (1993) made several modifications to the original procedure and recommended them for large datasets. However, it is not clear whether they provide utilities for small datasets. This study is to investigate their performances in small sample situations where tests with 20 or 40 items are administered to 30, 60, or 90 examinees. Item responses are generated so that half of the items measure one latent trait and the other half measure another, with their actual correlation being 1, 0.5, or 0. The original Stout's T and its modifications are obtained using SAS 9.1 and Microsoft Access 2003. Each of the 18 simulated conditions is replicated 100 times and the observed error rates are obtained to illustrate the performances of the modified procedures.

**Rijmen, F.**  
Educational Testing Service, Princeton, NJ (frijmen@ets.org)

IS-Tuesday-am-s25

### Beyond HMMs: Full Information Maximum Likelihood Estimation in Limited Time for a General Class of Latent Variable Models

For categorical data, the use of models with a multidimensional latent structure has been hampered because the number of computations involved in brute force integration over the latent space is exponential in the dimensionality of the model. The widespread use of the hidden Markov model is partly due to the fact that an efficient estimation algorithm is available for this model. The algorithm exploits the conditional independence relations between the latent variables. It is shown how this algorithm is a special case of a very general method. Specifically, the statistical model is associated with a graph in which nodes correspond to random variables and the edges represent conditional dependency relations between the variables. The core of the method consists of applying transformations to this initial graph. The structure of the transformed graph provides a factorization of the joint probability function, which is the basis of a modified and more efficient E-step of the EM algorithm. The method is applicable to a large family of models, and proceeds in a fully algorithmic way. Hence, it overcomes the computational burden that has (too) long been clouded the sky for latent variable models for categorical data.

**Roberts, J.S.**, and Thompson, V.M.  
School of Psychology, Atlanta, GA (james.roberts@psych.gatech.edu)

IS-Monday-pm-s17

### Accuracy of Alternative Parameter Estimation Methods with the Generalized Graded Unfolding Model

The generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000) is a unidimensional item response theory model for unfolding responses to traditional Likert or Thurstone style questionnaires. Parameters in the GGUM have historically been estimated using a marginal maximum likelihood (MML) approach which generally yields accurate estimates with reasonable data demands when polytomous item responses are studied (Roberts, Donoghue & Laughlin, 2002). However, there is a systematic tendency for the MML method to overestimate item locations and corresponding subjective response category thresholds for extreme items when binary responses are used (Cui, Roberts, & Bao, 2004). This paper describes a simulation study that compares and contrasts the accuracy of the MML approach with a marginal maximum *a posteriori* (MMAP) technique and a Markov Chain Monte Carlo (MCMC) method. The simulation examines the effects of sample size, test length, number of item response categories, and estimation method on the accuracy of both item and person parameters in

## ABSTRACTS OF THE CONTRIBUTIONS

---

the GGUM. Particular attention is given to the case of binary item responses and the ability of the MMAP and MCMC methods to mitigate the bias in MML parameter estimates associated with extreme items.

**Rosopa, P.J.**, and Wolf, A.N.  
Clemson University, Clemson, SC (prosopa@clemson.edu)

Poster-Tuesday-pm-s50

On the Effects of Measurement Error on Statistical Inferences Based on Heteroscedasticity-Consistent Covariance Matrices

Heteroscedasticity in linear models can impair the efficiency of the usual least squares estimator, resulting in conservative or inflated Type I error rates (Box, 1954) as well as reduced statistical power (DeShon & Alexander, 1996). To deal with such models and if a variance-stabilizing transformation cannot be found, researchers have suggested various alternatives (Carroll & Ruppert, 1988; Fox, 1997) including statistical approximations for specific heteroscedastic linear models (see e.g., Alexander & Govern, 1994; Luh & Guo, 2002). White (1980) proposed another alternative, a heteroscedasticity-consistent covariance matrix. Two variants—HC3 (Mackinnon & White, 1985) and HC4 (Cribaro-Neto, 2004)—will be investigated because they are considered robust, the pattern of heteroscedasticity can be of an unknown form, and general linear hypothesis tests can be conducted as usual (see e.g., Cribaro-Neto, 2004; Long & Ervin, 2000). Interestingly, the effects of measurement error on these techniques have not been investigated in previous research. This is an important issue because measurement error is of critical concern for researchers in many disciplines including psychology and allied fields (Nunnally & Bernstein, 1994). Using Monte Carlo methods, we consider HC3 and HC4 under various conditions including measurement error. Findings and implications for psychological research will be discussed.

**Roussos, L.**, and Ferdous, A.  
Measured Progress, Dover, NH (LRoussos@MeasuredProgress.org)

TS-Wednesday-am-s66

A Validity Study Comparing the Results of Skills Diagnosis and Standard Setting

Skills diagnosis models have been developed to carry out mastery classification of examinees using sophisticated statistical methods. There is a pressing need to validate these models with real data. Because mastery classification is closely related to classification of students with respect to state standards, the current paper compares the results of standard setting with the statistical results from skills diagnosis models. The data come from two standard settings that have recently been conducted, one using a Bookmark method and one using a yes/no variation of Angoff method. The Angoff standard setting was conducted to set grade 10 math cut scores. The test had 105 items and included data from over 4000 students. The Bookmark standard setting was conducted to set grade 3 reading cut scores. This test had 34 items and included data from over 10,000 students. The paper develops several methods for explicitly connecting the components of the skills diagnosis models with the cut-score judgments of the expert panelists who perform the standard setting. Furthermore, a method is developed for translating the skills diagnosis results into cut-scores. Preliminary results show a moderate relationship between the cut scores from the standard setting and those implied by the skills diagnosis statistical results.

**Rust, J.**, Golombok, S., Hines, M., Zervoulis, K., Croudace, T., and Golding, J.  
University of Cambridge, Cambridge, United Kingdom (jnr24@cam.ac.uk)

IS-Monday-pm-s16

Developmental Trajectories of Sex-Typed Behavior in Boys and Girls

The stability of sex-typed behavior from the preschool to the middle school years was examined in a UK general population birth cohort sample (ALSPAC – the Avon Longitudinal Study of Parents and Children). The Pre-School Activities Inventory (PASI) a measure of within-sex variation in sex-typed behavior, was completed by the primary caregiver when the child was 2½, 3½ and 5 years, and a modified version, the Child Activities Inventory, was completed by the child at age 8. The investigation involved a general population sample of 2726 boys and 2775 girls.

## ABSTRACTS OF THE CONTRIBUTIONS

---

Linear mixed models were applied using latent growth modeling. Sex-typed behavior increased through the pre-school years and those children who were the most sex-typed at age 2½ were still the most sex-typed at age 5, with those children who showed the highest levels of sex-typed behavior during the pre-school years continuing to do so at age 8.

**Ryoo, J.**, and Seo, D.G.

CS-Tuesday-am-s30

University of Minnesota, Minneapolis, MN (ryoox001@umn.edu)

### Efficiency of Bayesian Estimation with Gibbs Sampling in IRT Model

Markov chain Monte Carlo (MCMC) methods have been recently used on item response theory (IRT) for parameter estimation. For a finite sample, MCMC methods have been developed but the numerical computations using MCMC are complex and time consuming. Classical IRT estimation methods such as maximum likelihood estimate (MLE), expected a posteriori (EAP), maximum a posteriori (MAP) have been used for person parameter estimation after calibrating item parameters based on marginal maximum likelihood (MML) in item response models. This study compared MML method for item parameter estimation and MLE, EAP and MAP methods for  $\theta$  estimation with MCMC estimation method using Gibbs sampling implemented by WinBUGS (Lunn, Thomas and Spiegelhalter, 2000). This study mainly focused on the three-parameter logistic model (3PLM). Through simulation studies, the parameter recovery of estimated item parameters and person parameters were evaluated for MML, MLE, EAP and MAP and MCMC method. Results indicated that the item and person parameter estimates using MCMC method with Gibbs Sampling are more accurate than MML, MLE, EAP and MAP methods based on root mean square error (RMSE) and correlation of true and estimated  $\theta$ s. In addition, the correlations between true and estimated  $\theta$ s showed that the capability of parameter recovery was improved as the number of items increased for both classical IRT estimation methods and MCMC method with Gibbs sampling.

**Samejima, F.**

CS-Monday-pm-s20

University of Tennessee, Knoxville, TN (fsamejim@utk.edu)

### Some Constancy in Amount of Item Information and its Loss Caused by Noise

Samejima (1982) has noticed there is some type of constancy in the amount of item information of a binary item, that is, when the item characteristic function (ICF) is strictly increasing in the latent trait  $\theta$  with zero and unity as its two asymptotes, the area under the square root of item information function (global item information) equals  $\pi$  regardless of different mathematical formulas of the ICF in the models, and across different value(s) of item parameter(s) within a single model. In the present paper, it will be demonstrated how non-zero lower asymptote and/or non-unity upper asymptote of the ICF affect the global item information, that is, they make the global item information decrease from  $\pi$ , and the amount of global item information is determined solely by the values of two asymptotes. The usefulness of this discovery, together with the local amount of item information, given  $\theta$ , will be discussed. For example, in computerized adaptive testing (CAT), there is no reason not to mix items that follow different mathematical models. Depending on the purpose of CAT, optimal mixture of items that follow different models may be considered.

**Sano, M.**

CS-Wednesday-am-s58

Prometric Japan Co., Ltd., Tokyo, Japan (makoto.sano@prometric.com)

### Detecting Overestimation of Slope Parameter Under Surface Local Dependence

Local item independence is generally a strong assumption for applying item response theory. Chen and Thissen (1997) proposed two models of local item dependence. One of the two is surface local dependence (SLD) which typically affects overestimations of slope parameters. This study evaluates the performance of some local item dependence indices focusing on the SLD (and item chaining condition). Simulation study was performed and

## ABSTRACTS OF THE CONTRIBUTIONS

---

computed local dependence indices with jIRTNew (Tsai Hsu, 2005). The approach of using information entropy is promising for detecting SLD and overestimation of slope parameter rather than applying X2 or G2 indices. Also we tried to apply two-dimensional full information item factor analysis with TESTFACT (SSI, 1998) for detecting overestimations of slopes. In some cases, TESTFACT successfully detects SLD as the second dimension of the slopes but it does not resolve the overestimations of the slopes. The study suggests that the local item dependence indices applying information entropy are superior for detecting SLD and overestimations of slope parameters.

**Satorra, A.**

CS-Monday-am-s9

Universitat Pompeu Fabra, Spain (albert.satorra@upf.edu)

### A $\chi^2$ Goodness-of-Fit Test Statistic for the Exploratory Factor Model

The exploratory factor model is often used in the practice of multivariate analysis and it is estimated frequently using methods different than ML. A  $\chi^2$  goodness-of-fit testing may be helpful in determining the number of factors to retain. A Wald-type chi-square goodness-of-fit test that can be used in conjunction with classical estimation methods different than ML is presented. Theoretical and computational issues of this test statistics are discussed. Simulated data is used to illustrate the issues discussed.

**Scholl, L.H.,** and Ye, F.

CS-Tuesday-pm-s45

University of Pittsburgh, Pittsburgh, PA (lhs1@pitt.edu)

### The Impact of Inappropriate Modeling of Cross-Classified Data Structures on Random-Slope Models

Cross-classified data structures are prevalent in social science research. For example, students are nested within middle and high schools, and the middle and high schools are crossed because not all students who attended a particular middle school attended the same high school. However, due to the complexity of cross-classified models, many researchers are hesitant to use these models to analyze such datasets. Instead, researchers often utilize strictly hierarchical models even though they are not reflective of the true data structure. Therefore, in order to assess the effects of using a model that ignores the cross-classified data structure, a simulation study was conducted. In this study, a strictly hierarchical model and a cross-classified model were both used to separately analyze a cross-classified dataset with random intercepts and slopes for both cross-classified factors. The results were compared through RMSE values and various measures of bias for the parameter estimations and the standard errors. The simulated dataset included a level-1 predictor and a predictor for each crossed-classified factor. The intraclass correlation, the correlations between the level-2 residuals, the number of cross-classified factors, and the mean size of the first cross-classified factor were manipulated.

**Seo, M.,** and Roussos, L.

CS-Tuesday-am-s36

University of Illinois at Urbana-Champaign, Sterling Heights, MI  
(minseo@uiuc.edu)

### Formulation of a Dimtest-Based Effect Size Measure (DESM) and Evaluation of DESM Estimator Bias

The purpose of the current study is to propose an interpretable effect size measure for use with DIMTEST, a nonparametric statistical procedure for testing the hypothesis of test unidimensionality. Having an effect-size measure greatly improves the utility of DIMTEST so that one can decide in the case of statistical rejection whether the amount of multidimensionality is enough to be of practical concern. After formulating the DIMTEST-based Effect Size Measure (DESM), the study evaluates the efficacy of DESM by comparing the DESM estimator with the DESM population parameter derived through integral formula using the Reckase and Mckinley (1991)'s multidimensional item response theory model. The simulation study was conducted to prove that the DESM estimator converges to the DESM parameter as well as to evaluate the amount of statistical bias present in the DESM estimator according to sample size, test length, and correlations between dimensions under two cases;

## ABSTRACTS OF THE CONTRIBUTIONS

---

Unidimensional case and two-dimensional case. To eliminate estimation error due to random error, the simulation study was replicated 400 times independently. The results indicated that the DESM estimator converges to its parameter but exhibits statistical bias. For the simulated test that had longer test length and bigger PT size, the bias was smaller.

**Serrano, D.**, and Stucky, B.D.

University of North Carolina at Chapel Hill, Chapel Hill, NC  
(dserrano@email.unc.edu)

CS-Monday-am-s8

### Response Time IRT

In this talk we present developments on a general class of item response models for response time data, which are shown to be best fit by time-to-event or survival distributions. Statistical theory, simulation results, and empirical applications are reviewed. Response time data are routinely collected in psychological disciplines (e.g., cognitive, clinical, and developmental psychology along with psycho linguistics) Cognitive psychologists and psycho linguists use response times as indicators of cognitive and language processing respectively. Developmental and clinical psychologists use response times as indicators of psychopathology, notably autism and aggression. Common modeling techniques assume response time is normally distributed. An alternative approach conceives of response times as arising from a time-to-event or survival process where the event is defined as time of response. No psychometric method incorporating latent variables has been established for evaluating item properties or scoring item responses arising from time-to-event distributions. We propose an item response model where response times are modeled via time-to-event likelihoods. Likelihoods considered include the generalized Gamma, Weibull, Exponential, and Log-normal distributions. Estimation is based on a Quasi-Newton Marginal Maximum Likelihood algorithm with adaptive Gaussian quadrature. Preliminary results from simulation and empirical applications are presented. Discussion focuses on model parameterization and scoring.

**Setzer, J.C.**

James Madison University, Harrisonburg, VA (setzerjc@jmu.edu)

Poster-Tuesday-pm-s50

### Parameter Recovery of an Explanatory Modified Effort-Moderated Item Response Model

Explanatory item response theory models are becoming increasingly popular due to the flexibility provided by the nonlinear mixed model frameworks (De Boeck & Wilson, 2004). However, with respect to the data, we assume the item responses are obtained under effortful conditions. In low-stakes assessments, it is well known that effort-levels vary across both items and examinees (and their interactions; Wise, 2006; Wise & DeMars, 2005). The Effort-Moderated Item Response Model (EMIRM; Wise, & DeMars, 2006) was introduced to account for low-effort at the item level. The purpose of this study was to extend the EMIRM to a person-side explanatory model, while fixing the values of both the pseudo-guessing and discrimination parameters. The resulting model is an explanatory modified EMIRM (EMEMIRM). Moreover, I performed a parameter recovery study (using SAS) to examine the effects of sample size, test length, and amount of rapid-guessing on parameter estimation. Results indicate that, across all conditions, the EMEMIRM provided more reliable and less biased estimates of the fixed effects compared to a similar explanatory, non-effort moderated model. Increased sample sizes further reduced RMSE and bias estimates related to the EMEMIRM fixed effects. Other effects due to sample size, test length, and amount of rapid-guessing will be discussed.

**Sheng, Y.**

CS-Tuesday-am-s30

Southern Illinois University, Carbondale, IL (ysheng@siu.edu)

### An Investigation of IRT Models with Hierarchical Priors

Full Bayesian estimation procedures have been developed for the conventional IRT models, where item slope and intercept parameters can assume prior distributions with specified hyperparameters. Previous research indicates that this estimation procedure is sensitive to the choice of informative and noninformative priors for small samples. This is especially the case with the three - parameter model, as improper noninformative priors result in an undefined posterior distribution, which gives rise to unstable parameter estimates. Even with proper noninformative priors, the procedure either fails to converge or requires an enormous number of iterations for the Markov chain to reach convergence. This problem can be resolved by specifying the priors in a hierarchical fashion so that the hyperparameters for the item parameters are unknown and have their own prior distributions. The advantage of such hierarchical modeling is demonstrated by comparing its parameter recovery with that of the procedure where the hyperparameters are specified. Furthermore, the full Bayesian procedure for the IRT models with hierarchical priors is found to be not sensitive to the choice of the informative and noninformative priors for the hyperparameters of the item parameters. Hence, this procedure is recommended for the IRT models, especially in small sample situations.

**Shigemasu, K., and Okada, K.**

CS-Wednesday-am-s54

The University of Tokyo, Tokyo, Japan (kshige@bayes.c.u-tokyo.ac.jp)

### An Application of Bayesian Confirmatory Factor Analysis to Behavioral Genetics

A Bayesian approach is used to separate the covariance matrix into two parts--the genetic influence component and the environmental influence component, and then factor analyze each resulting covariance matrix. This model distribution does not assume the additivity of genes, and a new parameter is introduced to represent the degree of interaction of the genes involved. The prior distribution for the factor loading matrix is set by identifying the non-zero elements and the typical conjugate distribution is used for the remaining parameters. The MCMC method is applied to derive the information from the posterior distribution for the relevant parameters. The number of factors is determined by the Bayes Factor, and various degrees of its approximation and their performance was compared by using the simulated data and real data. Also, by means of the same technique, several patterns represented by the prior distributions were compared and evaluated theoretically.

**Shim, H.S., and Roberts, J.S.**

Poster-Tuesday-pm-s50

Georgia Institute of Technology, Atlanta, GA (hishin@gatech.edu)

### A New IRT Model to Estimate Differential Latent Change Trajectories in a Multi-Stage, Longitudinal Assessment

Repeated measures designs are widely used in educational and psychological research to compare the changes exhibited in response to a treatment. Traditionally, measures of change are found by calculating difference scores for each person. However, problems such as the reliability paradox and the meaning of change score arise from using simple difference scores to study change. The purpose of this study is to develop a new multidimensional item response theory (MIRT) model for repeated measurement analysis and to apply the new model to data from the Early Childhood Longitudinal Study - Kindergarten cohort (ECLS-K) to examine differential growth in math ability between male and female students from kindergarten through fifth grade. The new model is a three- parameter logistic model for longitudinal assessment, and multiple group structure is also incorporated into the model (e.g., males versus females over time). It allows for the estimation of the latent change scores instead of difference scores, addresses some of the limitations of using difference scores, and provides a direct comparison of the mean latent changes exhibited by different groups. A simulation-based test was conducted to test the viability of the model and results indicate the newly developed model can estimate the parameters accurately.

**Shimizu, S.** <sup>(1)</sup>, Hoyer, P.O. <sup>(2)</sup>, and Hyvarinen, A. <sup>(2)</sup>

IL-Monday-pm-s13

<sup>(1)</sup> Osaka University, Osaka, Japan (shoheishimizu@mac.com)

<sup>(2)</sup> University of Helsinki, Helsinki, Finland

### Linear Non-Gaussian Structural Equation Models

Linear structural equation models (linear SEMs) are widely applied in many empirical studies including social sciences, neuroinformatics and bioinformatics. Estimation of linear SEMs for continuous variables typically uses covariance structure of data alone and poses serious identifiability problems so that many important models including path analysis models are indistinguishable with no prior knowledge on the structures. A linear acyclic model which is a special case of path analysis models is typically used to analyze causal influences. Covariance information alone is not sufficient to uniquely estimate such a linear acyclic model and in most cases cannot identify the full structure (path coefficients and directions) of the model. Bentler (1983, PMK) proposed that non-Gaussian structures of data could be useful to overcome such identifiability problems of covariance-based estimation of SEMs. Recently we showed that use of non-Gaussianity allows the full structure of a linear acyclic model to be identified without pre-specifying any path directions between the variables (Shimizu et al., 2006, JMLR). The new method is based on a fairly recent statistical technique called independent component analysis (Hyvarinen et al., 2001, Wiley). We will first present an overview of linear non-Gaussian SEMs and then go to some recent advances we have made.

**Shojima, K.**

CS-Tuesday-am-s29

The National Center for University Entrance Examinations, Tokyo, Japan  
(shojima@rd.dnc.ac.jp)

### Neural Test Theory: A Nonparametric Test Theory Using the Mechanism of a Self-Organizing Map

Neural test theory (NTT) is a nonparametric test theory that uses the mechanism of a self-organizing map. The latent scale assumed in the NTT is not continuous but rank-ordered, because a test does not have high enough reliability to continuously measure human abilities. That is, the most that a test can do is to sort examinees into several ranks. The NTT is a methodology for standardizing and equating tests provided that the latent scale is ordinal. The number of latent ranks is up to the teacher or test administrator, although it can also be statistically determined by using model-fit indices. The item reference profile (IRP) represents the expected correct answer ratios at respective latent ranks, and it is useful for reviewing the statistical characteristics of each item. The IRP does not always monotonically increase. However, a monotonic increase constraint can be imposed on the IRP in the estimation process. In addition, rank membership profile is the posterior distribution of each examinee's latent rank, and it can be used to determine the latent rank to which the examinee belongs. Furthermore, batch-type learning can be executed with an EM algorithm.

**Shrout, P.E.**, and Ledgerwood, A.

CS-Tuesday-pm-s47

New York University, New York, NY (pat.shrout@nyu.edu)

### Bias Reduction vs. Precision of Estimates in Mediation Analysis

Mediation analyses in the tradition of Baron and Kenny (1986) attempt to describe the extent to which the effect of an explanatory variable X on an outcome Y is carried through an intervening or mediating variable M. The analyst typically estimates the indirect (mediated) effect as the product of the structural coefficient (a) describing the effect of X on M and the structural coefficient (b) describing the effect of M on Y, adjusting for X. If measurement error in X and M is ignored, the product (a\*b) is biased and is typically too close to zero. When such bias is eliminated using a measurement model in the context of SEM, the analyst might expect greater statistical power to test the indirect effect because of a larger effect size. However, the relative imprecision of the SEM estimate is usually larger than the bias reduction, leading to less power and precision for the unbiased analysis. We describe this pattern using simulation studies that vary the mediation effect size and the amount of measurement error in X, M and Y. We

## ABSTRACTS OF THE CONTRIBUTIONS

---

recommend that mediation studies be planned with sample sizes that take into account adjustment for measurement error.

**Shu, L.**

TS-Wednesday-am-s59

CTB/McGraw-Hill, Monterey, CA (lianghua\_shu@ctb.com)

### A Comparison of Rasch Model Parameter Estimates between PARDUX and WINSTEPS

WinSteps (Linacre, 1991-2006) is a widely used program for Rasch model parameter estimation using joint maximum likelihood (JML). PARDUX (Burket, 1991-2005) estimates parameters simultaneously for dichotomous and polytomous items using marginal maximum likelihood (MML), via expectation-maximization (EM) algorithms (Bock & Aitkin, 1981; Thissen, 1982, Yen 1991). This proposal is to compare the recent versions of WinSteps (Linacre, 2006) with Pardux (Shu & Burket 2007) by using simulated data. Unlike JML, MML need pre-specify the trial  $\theta$  distribution. This is (0,1) normal distribution in Pardux. The questions are: does Pardux have an advantage/disadvantage over WinSteps for the data with normal/non-normal  $\theta$  distribution? Will sample size and test length affect on the accuracies of parameter and  $\theta$  estimations? The simulated responses have total 36 (=3X3X4) data sets, in which, there are short (10-item), medium (35-item) and long (60-item) tests with small, medium and large sample sizes and four different  $\theta$  distributions (normal, negatively/positively-skewed, and approximately symmetric but platykurtic overall). To compare the estimated and true item parameters and  $\theta$ s, it is necessary to place them on the same scale by using mean/mean equating procedure. The above proposed questions will be addressed through the comparisons.

**Shyu, C.-Y.**

Poster-Tuesday-pm-s50

National University of Tainan, Tainan city, Taiwan (cyshyu@mail.nutn.edu.tw)

### The Investigation of Impact of Scale-Transformation on Cross-Lingual Linking

The purposes of this study are to compare several procedures for scaling adapted item parameters, and to examine the effect of using these procedures on cross-lingual linking in the separate monolingual groups design. All procedures considered in the present study will be compared and evaluated via a simulation study. A Chinese version and English version of a national Science examination are separately administered to Chinese- and English- speaking examinees. The effect of using these procedures on cross-lingual linking will be examined. Three procedures will be implemented by first excluding DIF items identified by the SIBTEST or the logistic regression detection methods from the anchor items. Then, the Stoking-Lord method will be used to place the target language (TL) item parameters on the scale of the source language (SL) item parameters. Another procedure, matched groups calibration, place the TL and the SL items on the same scale without using anchor items. Then, IRT equating methods will be applied to get the TL form equivalent raw and scale scores. In the simulation study, four sets of the TL item parameters will be generated to reflect four types of change from the SL item parameters. In addition to examining the stability of the parameter estimates for the anchor items, three criteria will be used to evaluate the resulting equivalent scores at each score point and at overall scores. Two TL sample sizes of 500 and 2000 will be used.

**Sinharay, S., and Holland, P.W.**

IS-Tuesday-pm-s42

Educational Testing Service, Princeton, NJ (ssinharay@ets.org)

### Missing Data Assumptions of the Non-Equivalent Groups with Anchor Test Design and their Implications for Test Equating

The Non-Equivalent groups with Anchor Test (NEAT) design, which is widely used for equating tests, involves *missing data* that are *missing by design*. Three popular equating methods that can be used with a NEAT design are:

## ABSTRACTS OF THE CONTRIBUTIONS

---

the *poststratification equipercentile equating* method, the *chain equipercentile equating* method and the item response theory observed score equating method. These methods make distinct assumptions about the missing data in the NEAT design. These assumptions are usually neither explicit nor testable. There is a lack of studies examining the missing data assumptions and their implications on equating. This work is an attempt to fill that void. First, we describe the missing data assumptions for the three methods. We then compare the three methods using data sets from two operational tests. For each data set, we examine how the three equating methods perform when the missing data satisfy the assumptions made by only one of these equating methods. The collection of results from these analyses provide us with valuable information regarding the comparative performance of the three methods and promise to assist the equating practitioners in choosing the most appropriate method for equating.

**Skrondal, A.**

IS-Monday-am-s6

London School of Economics, London, United Kingdom (a.skrondal@lse.ac.uk)

### Prediction in Multilevel Generalized Linear Models

We discuss prediction of expected responses, such as predicted probabilities, in multilevel generalized linear models. Depending on the purpose, three different kinds of response expectations may be of interest: the marginal expectation, the conditional expectation, and the posterior expectation. We discuss how to obtain predictions of these expectations and illustrate their use.

**Smith, J., and Habing, B.**

TS-Wednesday-am-s59

University of South Carolina, Columbia, SC (smith.jessalyn@gmail.com)

### Alternative IRT Models: Incorporating Guessing Appropriately

This work investigates alternative item response theory models that are theoretically grounded and appropriately reflect the actual guessing behaviors observed in practice. The three-parameter logistic item response theory model accounts for differing item discrimination, difficulty, and a constant pseudo-guessing parameter for all examinees on a given item. However, it is intuitive that the effects of guessing vary by the ability level of the examinee. Studies (e.g., Walker & Liu, 2006 and Agnoff, 1989) have shown that lower ability examinees tend to guess worse than chance (e.g., the traditional constant value of the pseudo-guessing parameter). Specifically, it is thought that guessing may have a negative effect on examinees with partial knowledge (Walker & Liu, 2006). Examinees who have some partial knowledge may be more attracted to distracters, and therefore incorrectly respond to items when this happens. On the other hand, examinees with no partial knowledge may choose a response at random. Based on this logic, it is reasonable to assume that the item response function is no longer strictly increasing nor is there a constant chance of a correct guess given the subject does not know the correct answer. This presentation accesses specific alternative parameterizations of the traditional 3PL that incorporates the guessing phenomenon more accurately and attempts to estimate these models using Markov chain Monte Carlo.

**Smithson, M., and Verkuilen, J.**

CS-Tuesday-pm-s44

The Australian National University, Canberra, Australia  
(Michael.Smithson@anu.edu.au)

### Mixed Regression Models for Doubly Bounded Metric Data

Continuous dependent variables with scales bounded at both ends are fairly common throughout the human sciences. Several authors recently have advocated using the beta distribution to model such variables. A number of gaps remain in our knowledge about beta regression models. Chief among them is the absence of methods for handling dependent observations. In this paper we develop and explore generalized linear mixed models and related dependent-observation models for beta-distributed dependent variables. Explicit solutions of the maximum likelihood estimation equations are not generally possible for these models, but several numerical estimation

## ABSTRACTS OF THE CONTRIBUTIONS

---

techniques can be applied. We investigate two popular approaches: Marginal maximum likelihood (MML) and Bayesian Markov Chain Monte Carlo (MCMC). Examples include repeated-measures, autoregression, and multi-level designs.

**Song, J.**<sup>(1)</sup>, Walls, T.A.<sup>(1)</sup>, and Raïche, G.<sup>(2)</sup>

Poster-Tuesday-pm-s50

<sup>(1)</sup> University of Rhode Island, Kingston, Rhode Island (jaejoonsong@gmail.com)

<sup>(2)</sup> Université du Québec à Montréal, Montréal, Canada

### Non-Graphical Factor Extraction: Application to the Adolescent Smoking Consequences Questionnaire (ASCQ)

Raïche and colleagues (2008) compared various numerical solutions to the Cattell's scree test. According to Raïche and colleagues, numerical solutions either deal with the acceleration, or the formal scree part of the plot. The scree test Acceleration Factor and the Optimal Coordinate approaches can be used to develop more refined results for either Kaiser's rule or parallel analysis. In this poster, we test the non-graphical approaches for the Cattell's scree test in the Adolescent Smoking Consequences Questionnaire (Lewis-Esquerre, Rodrigue, & Kahler, 2005). The ASCQ is a 30-item measure of adolescents' expectancy towards smoking behavior. The ASCQ was measured for 122 adolescents (mean age = 18 years), as part of a three wave study of smoking behavior. Principal component analysis was conducted for wave 2 and wave 3 using nFACTORS package in statistical software R. The number of factors retained by the criterion of eigenvalues, parallel analysis, optimal coordinates and acceleration factor were evaluated. Numbers of factors retained were 9, 3, 3, 1 for wave 2 and 7, 3, 3, 1 for wave 3. This poster is an application of various non-graphical solutions for the Cattell's scree test. Findings using this technique for this scale relative to other techniques will be discussed.

**Sotaridona, L.S.**<sup>(1)</sup>, Long, D.<sup>(2)</sup>, and Park, S.-H.<sup>(2)</sup>

TS-Monday-pm-s22

<sup>(1)</sup> CTB/McGraw-Hill, Roanoke Rapids, NC (Leonardo\_Sotaridona@ctb.com)

<sup>(2)</sup> Tennessee Department of Education

### From Scannable Test Booklets to Separate Answer Sheets by Third Grade: Implications on Pre-Equating Design

Scoring efficiencies can lower administration costs and support faster score reporting. An efficiency considered by a state was to change the response mode from scannable test booklets to answer sheets for third-grade pupils. One popular belief for using scannable test booklet is that primary school age children are unable to use a separate answer sheet adequately. However, no research was available to provide guidance along state-wide assessments that used IRT in scoring. The purpose of the present study was to evaluate the impact of using a separate answer sheet on the quality of the item characteristics, performance classification, and pre-equating results using data sets from a state-wide assessment program. Two groups, similar in demographic composition, were compared; half of the subjects answered in the test booklets and half answered on separate answer sheets. Results between the two groups were virtually the same in terms of item parameter estimates, item misfits, omit rates, test reliabilities, performance levels, SEMs, and performance classifications. The results suggest that a switch from one response mode to another does not compromise pre-equating results.

**Steinley, D.**

SA-Monday-am-s2

University of Missouri-Columbia, Columbia, MO (steinleyd@missouri.edu)

### K-Means Clustering: State-of-the-Art Methodological Developments

The presentation synthesizes the results and methodology of current research conducted in the area of K-means clustering. The K-means method is introduced and several recommendations are provided for the stages of cluster

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

analytic decisions. Such decisions included variable standardization, variable selection, data reduction, and algorithmic initialization techniques.

**Stucky, B.D.**, and Thissen, D.

CS-Tuesday-am-s27

Department of Psychology, UNC-CH, Chapel Hill, NC (bstucky@email.unc.edu)

### Item Response Theory for Weighted Summed Scores

For tests constructed and evaluated using classical test theory (CTT), items have commonly been given differential weights, allowing them to contribute more or less to a composite measure. While item weighting has historically received much attention in CTT, rarely has a priori weighting been considered in an item response theory (IRT) framework (see Sykes & Hou, 2003 for an exception). IRT methods currently available for computing IRT scaled scores are limited to unit-weighted items. A new method of incorporating weights into IRT is proposed. This method could be implemented in situations where tests have explicitly chosen arbitrary weights that are applied prior to IRT analysis. Revising the recursive algorithm currently used to compute scaled score estimates for summed scores (Thissen, Pommerich, Billeaud, & Williams, 1995) enables scaled score estimates for weighted summed scores to be computed. This procedure allows the likelihoods of response patterns containing a given weighted summed score to be obtained yielding posterior distributions for both unit- and arbitrarily-weighted summed scores. An application of this procedure is used to discuss the relationship between arbitrary weights and discrimination parameters for expected a posteriors (EAPs) and their associated standard deviations.

Hwang, H., and **Suk, H.W.**

CS-Tuesday-am-s35

McGill University, Montreal, Canada (hye.suk@mail.mcgill.ca)

### Fuzzy Clusterwise Partial Least Squares Regression

Partial least squares regression (PLSR) is a multivariate regression method in which a set of orthogonal components are extracted from predictors in such a way that they explain the covariances between predictors and criteria as much as possible. This method is particularly attractive when the number of predictors is large compared to that of observations. However, PLSR is thus far designed for an aggregate-sample analysis based on the assumption that all observations come from a single homogenous population. Thus, it is not suitable for investigating whether there exist heterogeneous subgroups in the population, which involve distinct effects of predictors on criteria. In this paper, PLSR is combined with fuzzy clustering in a unified framework so as to account for this potential group-level heterogeneity. The proposed method can identify fuzzy clusters of observations and simultaneously estimate PLSR parameters within each cluster. Permutation tests are applied to test the significance of components. Cluster validity measures are used to decide the number of clusters. An empirical example is provided to illustrate the usefulness of the proposed method.

**Sun, R.**, and Willson, V.L.

CS-Wednesday-am-s63

Texas A&M University, College Station, TX (ronghua\_sun2001@tamu.edu)

### Misspecifications on Growth Mixture Modeling

The purpose of this study is to investigate the effects of growth mixture modeling misspecification in terms of model power and performance of fit statistics: LMR, BLRT and other criterion information. The effects of misspecifying a 2-Class Means-only Mixture model in four time-point latent growth model were the focus of this investigation. It was motivated by conditions in school growth studies in which students' entry skills may affect their rate of growth. Sample size (100, 200, and 500), class proportion (balanced and unbalanced design), Mahalanobis distance effect (0, 0.5 and 1.5), and effects of difference of the latent slope regressed on latent intercept (large and medium) were four factors studied, based in part on a data analysis by Muthén and Asparouhov (2003). Correctly specified models were

## ABSTRACTS OF THE CONTRIBUTIONS

---

examined to determine power and Type I error rates, and misspecified models were examined to evaluate the effects on power, Type I error rates, class enumeration and fit indices.

**Swartz, R.J.**<sup>(1)</sup>, Choi, S.W.<sup>(2)</sup>, and Herrick, R.C.<sup>(1)</sup>

TS-Monday-am-s11

<sup>(1)</sup> University of Texas, Houston, TX (rswartz@mdanderson.org)

<sup>(2)</sup> Evanston Northwestern Healthcare Research Institute & Northwestern University

### Computerized Adaptive Testing Item Selection Procedures for Dichotomous Items – What Do We Lose by Being Greedy?

Computerized adaptive testing has been shown to achieve better measurement efficiency and lower measurement error by adaptively selecting and administering items to estimate a person's unknown latent trait. Although a variety of item selection procedures have been proposed and are in use, most of them, including some Bayesian procedures, are myopic or greedy, i.e. they select the next best item given the current responses. In the statistical literature, under most circumstances such greedy algorithms are less than optimal; only in special cases will myopic algorithms be optimal. A fully Bayesian adaptive sequential design optimizes over all possibilities, (i.e. the possibility of asking subsequent items beyond the next one) not just the next one. However, this method is computationally intense and greedy algorithms are easier to implement. We provide simple hypothetical examples using small dichotomous item banks to compare a fully Bayesian adaptive sequential algorithm to several of the popular greedy algorithms, in terms of the characteristics of the items selected, root mean squared error of the theta estimate, and bias. We identify conditions when the fully sequentially adaptive item selection method is superior and when the greedy algorithm is "good enough."

**Takai, K.**, and Kano, Y.

CS-Monday-am-s10

Osaka University, Toyonaka, Japan (takai@sigmath.es.osaka-u.ac.jp)

### Test of Independence in a 2x2 Contingency Table with Nonignorable Nonresponse

In this talk, test of independence in a 2 x 2 contingency table with nonignorable nonresponse (NIN) missing data is discussed, where the likelihood ratio statistic is used. There are many models for the NIN missing mechanism, and most of them have the serious problem that some parameters are unidentified under the null hypothesis. The parameters must be specified before analysis. Simulations are conducted to investigate the sensitivity of various different choices of pre assigned values for such parameters to the statistic and compare our statistic with the one not taking the missing mechanism into account. It is found that the statistic considering the mechanism is more powerful. A real data analysis for crime data is also made.

**Takane, Y.**<sup>(1)</sup>, Hwang, H.<sup>(1)</sup>, and Abdi, H.<sup>(2)</sup>

CS-Tuesday-pm-s46

<sup>(1)</sup> McGill University, Montreal, Canada (takane@psych.mcgill.ca)

<sup>(2)</sup> University of Texas at Dallas, Dallas, TX

### Regularized Multiple-Set Canonical Correlation Analysis

Multiple-set canonical correlation analysis (Generalized CANO or GCANO for short) is an important technique because it subsumes a number of interesting multivariate data analysis techniques as special cases. More recently, it has also been recognized as an important technique for integrating information from multiple sources. In this paper we present a simple regularization technique for GCANO and demonstrate its usefulness. Regularization is deemed important as a way of supplementing insufficient data by prior knowledge, and/or of incorporating certain desirable properties in the estimates of parameters in the model. Implications of regularized GCANO for multiple correspondence analysis are also discussed. Examples are given to illustrate the use of the proposed technique.

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

**Takane, Y.**, and Oshima-Takane, Y.  
McGill University, Montreal, Canada (takane@psych.mcgill.ca)

IS-Tuesday-am-s26

### Nonlinear Multivariate Analysis via Artificial Neural Network Models

Many of the feed-forward artificial neural network models can be viewed as nonlinear extensions of traditional linear multivariate analysis techniques. In this talk we discuss two such techniques, multi-layer encoder (auto-associative) networks and multiple-source information integrator networks. The former is regarded as nonlinear principal component analysis, and the latter as nonlinear multiple-set canonical correlation analysis. We review recent developments in algorithm construction in these areas, and give examples of applications of these techniques.

**Tala, A.**  
Warsaw University, Warsaw, Poland (alexistala@yahoo.fr)

Poster-Tuesday-pm-s50

### Robust Interdisciplinary Measurement and Analysis of Wellbeing. Some Methodological Issues and Application

The paper aims to present a robust, interdisciplinary and comprehensive methodology of wellbeing measurement and analysis. We increase robustness and comparability by reducing researcher's subjective choices to the minimum and taking into account the cultural, social and economic contexts of the studied individuals and groups. The methodology is an iterative process involving stakeholders concerned by the problematic. It addresses the main questions of interest about wellbeing: its perception, constituents, characteristics, dimensions, assessment, determinants and consequences. The main steps of the methodology include: (1) systematic interdisciplinary research in the relevant literature; (2) consultation with resource people and stakeholders; (3) exploratory qualitative field research; (4) building a theoretical framework; (5) confirmatory qualitative and quantitative field research; (6) multidimensional exploratory and confirmatory data analysis with latent variable models (7) construction of wellbeing indicators for each dimension and for the overall construct; (8) validation of the constructs and methodology; (9) evaluating a causal model of the wellbeing with other socioeconomic variables and policy variables by Structural Equation Models (SEM), Partial Least Squares (PLS) and path analysis (PA). We will present some methodological issues concerning steps 1-5, and realize steps 6-9 with the data from some Polish databases.

**ten Holt, J.C.**, van Duijn, M.A.J., and Boomsma, A.  
University of Groningen, Groningen, The Netherlands (j.c.ten.holt@rug.nl)

Poster-Tuesday-pm-s50

### Construction and Evaluation in Practice: Factor Analysis Versus Item Response theory

In scale construction and evaluation, factor analysis (FA) and item response theory (IRT) are two methods frequently used to determine whether a set of items reliably measures a latent variable. In a review of 41 published studies it was examined which methodology (FA or IRT) was used, and what researchers' motivations were for applying either method. Characteristics of the studies were compared to gain more insight in the present practice of scale analysis. Findings indicate that FA is applied far more often than IRT. Many a time it is unclear whether the data justify the method because model assumptions are neglected. It is suggested that researchers (a) use substantive knowledge about the items to their advantage; and (b) investigate model assumptions and report corresponding findings, either in the paper or on a website.

**Thissen, D.**  
University of North Carolina, Chapel Hill, NC (dthissen@email.unc.edu)

IS-Wednesday-am-s53

### Discussion of the session "Item Response Theory with Nonnormal Latent Distributions"

## ABSTRACTS OF THE CONTRIBUTIONS

---

Cai, L., and **Thissen, D.**

CS-Wednesday-am-s64

University of North Carolina, Chapel Hill, NC (dthissen@email.unc.edu)

### Implementation of a Supplemented EM Algorithm to Compute the Error Covariance Matrix for ML Parameter Estimates in Item Response Theory Models

A *supplemented EM* (SEM) algorithm is a procedure that may be used to compute the error covariance matrix for maximum likelihood (ML) parameter estimates that are obtained with an EM algorithm. SEM algorithms use additional computation parallel to that of the main EM algorithm to compute a transformation that is applied to the complete-data information matrix to obtain the information matrix for the parameters given the observed data. SEM procedures are especially useful in contexts such as EM estimation for item response theory (IRT) models, in which the complete-data information matrix is already available as a component of the M-step computation. In addition to that complete-data information matrix, SEM procedures require only the “history” of the EM iteration and the E- and M-step implementations. However, there are a number of aspects of the generalized SEM algorithm that may be customized for the problem at hand. Prominent among these is the choice of a “window” of the EM history that is involved in the SEM computations; that choice and details of the convergence criterion greatly affect the procedure’s speed. This presentation discusses customization of the SEM algorithm for efficient implementation in the context of parameter estimation for commonly used IRT models.

**Thompson, N.A.**, and Ro, S.

TS-Wednesday-am-s66

Prometric, Saint-Paul, MN (Nathan.thompson@prometric.com)

### Item Exposure in Computerized Classification Testing

In computerized classification testing (CCT), or pass/fail adaptive testing, there are two types of item selection based on item response theory: maximizing information at the current ability estimate or at the cutscore. While selecting items at the cutscore can be more efficient (Spray & Reckase, 1994), it has greater item exposure effects, because all examinees will get the same sequence of items unless an exposure control is implemented. This is not the case with current estimate or “adaptive” item selection, where many different sequences are possible. It is important to better understand the interaction of termination criteria and item selection in CCT. While most comparisons of CCT methods are made in terms of average test length and classification accuracy, this study will compare methods based on item exposure rates, using monte carlo simulation. Three termination criteria will be evaluated, the sequential probability ratio test (Reckase, 1983), maximum likelihood ability confidence intervals (Weiss & Kingsbury, 1983), and a generalized likelihood ratio (Huang, 2004). This is in contrast to Kalohn and Spray (1998), who investigated only Bayesian confidence intervals; confidence intervals are the least efficient of the three termination criteria. Additional independent variables are item exposure constraints and item pool characteristics.

**Timmerman, M.E.**

SA-Wednesday-pm-s68

University of Groningen, Groningen, The Netherlands (m.e.timmerman@rug.nl)

### Principal Component Analysis and Generalizations

Principal Component Analysis (PCA) is a popular technique for condensing information in a large set of variables into a smaller set of components. Although the definition of Principal Components (PCs) is straightforward, the technique has several facets that are important to know of and to connect with each other. We will provide a thorough introduction of the technique and discuss issues important in applications, including the different criteria underlying PCA, rotation criteria and inferential aspects. In psychology, PCA is often used in test- and questionnaire construction to detect the structure in the variables. It has been debated contentiously whether PCA is an appropriate method in this context, or whether Common Factor Analysis (CFA) should be preferred. The essential differences between the two approaches will be explained, as well as their implications for the choice between PCA and CFA. Attention will be paid to frequently encountered misconceptions. Finally, we will provide an overview of the

## ABSTRACTS OF THE CONTRIBUTIONS

---

numerous generalizations of PCA and their applications. Examples of such generalizations are non-linear PCA, Redundancy Analysis, Common Principal Components, Multilevel Component analysis and Three-way Component Analysis.

**Tofighi, D.**<sup>(1)</sup>, MacKinnon, D.P.<sup>(1)</sup>, and Yoon, M.<sup>(2)</sup>

CS-Tuesday-pm-s47

<sup>(1)</sup> Arizona State University, Tempe, AZ (dtofighi@asu.edu)

<sup>(2)</sup> Texas A&M University

### Covariance Among Regression Coefficients Estimates in a Single Mediator Model

Mediation models have been the focus of much substantive and statistical research. This study presents formulae for the covariances between parameter estimates in a single mediator model. These covariances are necessary to derive the multivariate-delta method standard errors, which in turn, are used to build confidence intervals for effect size measures in a single mediator model. Effect size measures are vital as they provide a meaningful way of comparing the mediated effect across mediation studies regardless of sample sizes. Out of six covariances in a single mediator model, only three have been proposed in the literature. In this study, we first analytically derived the covariances between all of the parameter estimates in a single mediator model. Using the derived covariances, we computed the multivariate-delta standard errors, and built the 95% confidence intervals for the effect size measures. A simulation study evaluated the accuracy of the standard errors in terms of relative bias and root mean squared errors (*RMSE*). In addition, we empirically evaluated the Type I error and power of the confidence intervals using various parameter values and sample sizes. Finally, we presented a numerical example and a SAS MACRO that calculates the confidence intervals for the effect size measures. This research was supported by the National Institute on Drug Abuse grant DA09757.

**Trierweiler, T.J.**

TS-Tuesday-am-s31

Fordham University, Bronx, NY (tjtrier@gmail.com)

### An Empirical Examination of Current Reporting Techniques in Applied Structural Equation Modeling Research

Research reviewing the applications of structural equation modeling (SEM) in applied social science research has called to attention several chronic problems regarding the uses and reporting of these techniques (MacCallum and Austin, 2000; Schreiber, Stage, King, Nora, & Barlow, 2006). Several researchers have proposed guidelines and recommendations outlining the essential information to be included in studies using SEM as the primary method of statistical analysis (see MacCallum & Austin, 2000, Schreiber et al., 2006; Raykov, 1991). No recent studies however have critically examined whether researchers are currently following these recommendations when reporting SEM analyses in the social and psychological sciences. The study reported here will attempt to address this concern by critically reviewing the reporting techniques of researchers using applied SEM techniques as their primary statistical analysis in peer reviewed journal articles published between 1998 and 2007. A random sample of articles ( $N = 600$ ) were reviewed and coded according to recommended reporting guidelines found in the literature. Studies were compiled and classificatory and descriptive analyses were conducted. Results suggest that very few researchers are following the guidelines recommended in the literature. Reasons for this and recommendations for strengthening these guidelines will be discussed.

**Tsai, R.C.**

Poster-Tuesday-pm-s50

National Taiwan Normal University, Taipei (rtsai@math.ntnu.edu.tw)

### Random Utility Models for Approval Voting

The method of approval voting is a commonly used voting procedure. In an approval voting task, each judge is usually asked to select a subset of the choice alternatives. In this poster, we introduce a Thurstonian-based random

## ABSTRACTS OF THE CONTRIBUTIONS

---

utility framework to investigate the underlying mechanism undertaken in the choice process and therefore provide a means to understanding the voting behavior. By postulating a more general formulation for the approval voting mechanism, we are able to extend previous approach to analyze data with the number of alternatives greater than three. More importantly, the “none” and/or “all” approval patterns were commonly excluded in the previous studies which attempt to model the underlying decision mechanism under approval voting or the “choose any” tasks. Here we propose models which allow for the possibility of “none” and/or “all” and therefore attempt to provide a comprehensive account for all the response patterns in the choice behavior. Based on the analysis of some empirical data, we found that the Thurstonian-based random utility models, unlike in the case for three alternatives, do not seem to conform to the true underlying mechanism or individual difference in the data with four choice alternatives.

**Tuerlinckx, F.**

IS-Wednesday-am-s61

K.U. Leuven, Belgium (francis.tuerlinckx@psy.kuleuven.be)

### Some Disturbing Facts about the Relation Between IRT and Response Times

The dominant variable of analysis in IRT is a single discrete response (often binary, sometimes polytomous). Unfortunately, other possibly valuable sources of information, such as response times, are often not considered while they may provide additional information about the underlying response process and lead to better inferences. In this talk, we present some findings from answers to personality items under speed conditions. It is found that the response times are related to the discrimination parameter. In addition, it will be explained why this relation between response times and the discrimination parameter is disturbing by making a link to a popular model for speeded decision making, the diffusion model.

**Uher, R.,** Perroud, N., Aitchison, K.J., McGuffin, P.,  
and The GENDEP Consortium

IS-Monday-pm-s16

King's College, London, United Kingdom (rudolf.uher@iop.kcl.ac.uk)

### Characterisation of Response to Antidepressants in a Pharmacogenetic Study

The aim of pharmacogenetics is to identify genomic predictors that explain the substantial individual variability in positive and adverse effects of medication. This is especially important for antidepressant medications, as their efficacy in large proportion of individuals with depression has been questioned and risk of severe adverse effects including treatment-emergent suicidality has been highlighted. Accurate definition of the response to antidepressant medication is needed to optimize the pharmacogenetic testing. In the Genome Based Therapeutics for Depression (GENDEP) study, over 800 adults with major depression have been treated with one of two antidepressant drugs. Depressive symptoms have been measured on 14 occasions in weekly intervals, using three established scales: Montgomery-Asberg Depression Rating Scale, the Hamilton Rating Scale for Depression, and the Beck Depression Inventory. We have applied modeling with latent variables to integrate the 3 measures of depression, explore heterogeneity of depressive symptoms, examine the shape of individual trajectories of symptom change over time, established individual differences in the delay of therapeutic response and relate symptom change to dose titration. The detailed description of symptom change over time has allowed differentiation of the specific effects of individual antidepressants that was not apparent with the original scales, and will inform the pharmacogenetic analyses.

**Umehara, T.,** and Kano, Y.

CS-Wednesday-am-s58

Division of Mathematical Science, Osaka University, Japan  
(umehara@sigmath.es.osaka-u.ac.jp)

### Support Vector Machine for the Dataset Including Multivariate Discrete Variables

## ABSTRACTS OF THE CONTRIBUTIONS

---

In IMPS2007, we proposed a new method for Support Vector Machine which can deal with mixture of continuous and discrete variables. In the method, we assigned a discrete variable to a proper value to minimize test error. We showed by numerical experiments that use of Radius Margin Bound minimizes test error. In this presentation, we apply this method to datasets including multivariate discrete variables. And we propose how to analyse such a dataset and show its performance by numerical experiments.

**Unkel, S.**, and Trendafilov, T.

CS-Wednesday-am-s55

The Open University, Milton Keynes, United Kingdom (s.unkel@open.ac.uk)

### Simultaneous Parameter Estimation in Exploratory Factor Analysis by Weighted Least Squares

The standard fitting problem in Exploratory Factor Analysis (EFA) is to find estimates for the factor-pattern matrix and the matrix of unique-factor variances which give the best fit to the sample correlation matrix with respect to some goodness-of-fit criterion. Predicted factor scores can then be obtained as a function of these estimates. Unlike factoring a correlation matrix, fitting the EFA model directly to the data matrix yield simultaneous solutions for both loadings and factor scores. In this paper, a new approach to the simultaneous estimation of all EFA parameters is considered. The EFA model is fitted to the data by minimizing a weighted least squares (WLS) loss function. The WLS fitting problem is solved by iteratively performing steps of an existing algorithm for unweighted least squares fitting of the same model. The approach is based on minimizing an auxiliary function that majorizes the WLS loss function. Numerical examples illustrate the performance of the proposed approach. This research was originated when considering a noisy version of Independent Component Analysis (ICA) based on the EFA model. For the needs of ICA one has to employ factoring the data matrix rather than its covariance/correlation matrix.

**Usami, S.**

Poster-Tuesday-pm-s50

University of Tokyo, Tokyo, Japan (joker1984927@yahoo.co.jp)

### Generalized Graded Unfolding Model with Manifest Variables

The generalized graded unfolding model (GGUM) has attracted increasing interest in item response theory (IRT). GGUM is capable of analyzing polytomous scored, unfolding data such as agree-disagree responses to attitude statements. When we desire to see the relation between a latent trait  $\theta$  and other variables (e.g. age and sex of examinees), regression analysis and correlation analysis are often performed after estimating each examinee's latent trait  $\theta$ . In this way, however, the estimates of regression coefficient and correlation coefficient are bound to be attenuated since the estimates of  $\theta$  are contaminated with errors. In the present study, the author proposed an item response model in which manifest variables are included in order to avoid attenuation. To compare the accuracy of estimates between the proposed model and GGUM (with no manifest variables in the model), the author performed MCMC simulation with different numbers of items and examinees. As a result, the proposed model consistently showed more accurate estimates. This research concludes with verifying the efficacy of the proposed model, using a real data of attitude measurement.

**van der Ark, L.A.**, and van der Palm, D.W.

CS-Wednesday-am-s62

Tilburg University, Tilburg, The Netherlands (a.vdark@uvt.nl)

### A New Reliability Coefficient Based on Latent Class Analysis

A reliability coefficient is proposed based on the unrestricted latent class model. The coefficient is a direct estimate of the reliability and not a lower bound such as Cronbach's alpha or Guttman's lambda. The coefficient is a refinement of reliability coefficients  $\rho_1$  and  $\rho_2$  which were proposed in the context of nonparametric item response theory. For the new coefficient, a latent class model is used to estimate the theoretical probability that a respondent obtains the same item score twice when an item is administered twice under identical conditions. If the number of latent classes is chosen sufficiently large, then this probability will be estimated accurately. In practice

## ABSTRACTS OF THE CONTRIBUTIONS

---

only a limited number of latent classes can be estimated and computation time increases rapidly with the number of latent classes. We studied the bias and computation time of the new coefficient under several conditions with respect to numbers of items and numbers of latent classes and compared them with bias and computation of existing reliability coefficients. Tentative results indicate that the new coefficient has less bias than other reliability coefficients ( $\alpha$ ,  $\lambda$ ,  $\rho_1$ ,  $\rho_2$ ) even when a latent class model is used with a limited number of latent classes.

**van der Leeden, R.**, Pieper, S., Brosschot, J.F., and Thayer, J.F.  
University of Leiden, The Netherlands (vanderleeden@fsw.leidenuniv.nl)

Poster-Tuesday-pm-s50

### Multilevel Analysis of Daily Process Data: Immediate and Prolonged Cardiac Effects of Momentary Assessed Stressful Events and Worry Episodes

Multilevel regression modelling has become a widely used method to analyze data obtained in clinical research, especially concerning longitudinal research designs. Research questions concern a variety of topics, such as the treatment of depression, hypochondriacal complaints, deliberate self-harming behavior, suicidal behavior, daily processes and so on. Research designs most often concern a mixture of both within and between factors. For instance, subjects are randomly assigned to a few conditions (treatment groups), which are repeatedly measured over time, or subjects from different populations are requested to record several variables on a daily basis. For this poster presentation we focus upon the multilevel analysis approach to daily process data, a type of longitudinal data for which the dependent variable is not modeled as a function of the time variable, but as a function of other predictors. To illustrate this approach we discuss research results from a study of the immediate and prolonged effects of stress and worry on cardiovascular activity, in particular heart rate and heart rate variability.

**van der Linden, W.J.**  
CTB/McGraw-Hill, Monterey, CA (wim\_vanderlinden@ctb.com)

IS-Tuesday-am-s32

### Local Observed-Score Equating

Each of the current methods of observed-score equating uses a single equating transformation for a population of test takers. But because test takers with different abilities have different observed-score distributions, these transformations always have to compromise between the individual score distributions. Hence, their equated scores are necessarily biased. The only way to reduce the bias is to replace the single transformation by a family of transformations and pick the member from this family that is locally best for each individual test taker. This idea of local equating will be explored for several observed-score equating problems, including linear equating, common item equating, and IRT observed-score equating. Results from empirical studies will be presented that confirm our expectation of a large bias for the conventional equating methods and help us to identify local methods with minimum bias. Finally, it will be shown how the framework of local equating helps us to address the as yet intractable problem of equating observed score on tests that are differentially speeded.

**van der Linden, W.J.**  
CTB/McGraw-Hill, Monterey, CA (wim\_vanderlinden@ctb.com)

IS-Tuesday-am-s61

### Conceptual Issues in Response-Time Modeling

Two different traditions of response-time (RT) modeling are reviewed: (i) the tradition of distinct models for RTs and responses and (ii) the tradition of model integration in which RTs are incorporated in response models or the other way around. Several conceptual issues underlying both traditions are made explicit and analyzed for their consequences. We then propose a hierarchical modeling framework consistent with the first tradition but with the integration of their parameter structures as a second-level of modeling. Two examples of the framework are presented. Also, a fundamental equation is derived which relates the RTs on test items to the speed of the test taker

## ABSTRACTS OF THE CONTRIBUTIONS

---

and the labor intensity of the items. The equation serves as the core of the RT model in the framework. Finally, the flexibility of the framework in predicting empirical correlations between RTs and responses is demonstrated.

**van der Maas, H.L.J.** <sup>(1)</sup>, Straatemeier, M. <sup>(1)</sup>, Klinkenberg, S. <sup>(1)</sup>, and Maris, G. <sup>(2)</sup> IS-Wednesday-am-s61

<sup>(1)</sup> University of Amsterdam, Amsterdam, The Netherlands

(H.L.J.vanderMaas@uva.nl)

<sup>(2)</sup> Cito, Arnhem and University of Amsterdam, The Netherlands

### Using RT in Adaptive Testing

We developed a child monitoring system to follow the development of arithmetic abilities of children during the primary school period based on weekly measurements. This system is based on three new psychometric innovations a) a scoring rule to incorporate RT in scoring b) using RT's in the adaptive test procedure to allow for expected probabilities correct much higher than .5 c) online estimation of item ratings so that pre-testing is not required. In this talk I will focus on the first two points, discuss the set-up of our system and present empirical results.

**Van Ginkel, J.**

CS-Monday-am-s10

Leiden University, Leiden, The Netherlands (jginkel@fsw.leidenuniv.nl)

### Further Investigation of the Influence of Simple Multiple-Imputation Methods on Psychometrically Important Statistics

Two-way imputation with factor loadings and two-way imputation for separate scales are multidimensional extensions of two-way imputation with error, a simple method for handling missing data in questionnaire data. Earlier research has shown that these methods produce small bias in psychometrically important statistics. In practice, researchers may be faced with situations that are more complex than the circumstances under which these methods were originally studied. For example, items may be contraindicative or may have unacceptably low factor loadings on every subscale, or complex missingness patterns may complicate computations. Simulations are carried out to study how robust methods two-way with factor loadings and two-way for separate scales are to complex situations, and are compared with methods two-way with error and multivariate normal imputation. The results show that both methods produce small bias in psychometrically important statistics in complex situations, comparable to the bias of multivariate normal imputation.

**Van Mechelen, I.**

IS-Tuesday-am-s26

K.U. Leuven, Belgium, (iven.vanmechelen@psy.kuleuven.be)

### Mapping the S-O-R Structure that Characterizes the Contextualized Personality of a Single Individual

Within contextualized personality research, one may wish to capture the networks of relations between situational features, cognitions/affects, and behaviors that characterize the behavioral signatures of individuals, along with their underlying process dynamics. In this paper, I explain how this goal can be achieved for a single individual, making use of a novel variant of hierarchical classes modeling of presence/absence information of a set of cognitions/affects and a set of behaviors in a set of situations. The modeling induces a set of relevant situational features from the data, along with if-then type relations between these features and cognitions/affects, and between cognitions/affects and final behaviors. The whole can further be displayed making use of a novel, insightful graphical representation. We illustrate with cognitive/affective and behavioral data from an eating-disordered patient.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Van Mechelen, I.**<sup>(1)</sup>, and Junker, B.W.<sup>(2)</sup>

CS-Monday-pm-s21

<sup>(1)</sup> K.U. Leuven, Belgium (iven.vanmechelen@psy.kuleuven.be)

<sup>(2)</sup> Carnegie Mellon University, Pittsburgh, PA

### A Skill Model for Cognitive Diagnosis Research

Given information on the requisite relation among a set of items, one may wish to induce from this information the underlying skills as involved in each item. This problem is shown to come down to a conjunctive skill factorization of the requisite relation under study. The factorization is further shown to imply a novel one-mode hierarchical classes model for two-way one-mode binary data. Existence and uniqueness of a skill factorization for a requisite data matrix at hand is discussed, along with algorithmic principles to arrive at exact as well as approximate skill factorizations. The approach is shown to be useful both for cognitive diagnosis research and for the construction of approximate, insightful representations of data on order relations of a type other than a requisite one, such as taxonomic (kind of) and meronomic (part of) relations.

**van Rosmalen, J.**, Koning, A.J., and Groenen, P.J.F.

IS-Tuesday-pm-s43

Erasmus University, Rotterdam, The Netherlands (vanrosmalen@few.eur.nl)

### Optimal Scaling of Interaction Effects in Generalized Linear Models

Multiplicative interaction models, such as Goodman's RC(M) association models, can be a useful tool for analyzing the content of interaction effects. However, most models for interaction effects are only suitable for data sets with two or three predictor variables. In this presentation, we discuss an optimal scaling model for analyzing the content of interaction effects in generalized linear models with any number of categorical predictor variables. This model, which we call the optimal scaling of interactions (OSI) model, is a parsimonious, one-dimensional multiplicative interaction model. We discuss how the model can be used to visually interpret the interaction effects. Several extensions of the one-dimensional model are also explored. Finally, an empirical data set is used to show how the results of the model can be applied and interpreted.

**van Schuur, W.H.**

IS-Monday-pm-s17

University of Groningen, Groningen, The Netherlands (H.van.Schuur@rug.nl)

### The Circumplex: An Ordinal Circular Unfolding Model for Polytomous Data

Some models for the measurement of attitudes, emotions, values, or personality assume a circular latent trait, called a circumplex (e.g., Plutchik & Conte 1997, Schwartz & Bilsky 1987, Holland 1985). A nonparametric confirmatory test of the circumplex is proposed, as well as an exploratory bottom-up hierarchical clustering procedure to find a maximal subset of items that forms a circumplex. The proposed model generalizes the model for dichotomous data (Mokken, Van Schuur, Leeferink 2000), which was based on the ordinal IRT model for dominance data (Sijtsma, Debets & Molenaar 1990) and proximity data (Van Schuur 1992) to polytomous data. An ordinal circular unfolding model fits the data perfectly if for every ordered quadruple of items the response patterns “low-high-low-high” or “high-low-high-low” are absent. Homogeneity of a circular unfolding scale is defined in terms of Loevinger's (1948) coefficient of homogeneity:  $1 - E(\text{obs})/E(\text{exp})$ , where  $E(\text{obs})$  is the number of ordered quadruples that violate the circular unfolding model, and  $E(\text{exp})$  the expected number under statistical independence. Diagnostic matrices for a probabilistic circumplex model are also presented.

**Verkuilen, J.**

TS-Wednesday-am-s59

CUNY-Grad Center, New York, NY (jverkuilen@gc.cuny.edu)

### Principles for Specification and Parameterization of Ideal Point Item Response Models

## ABSTRACTS OF THE CONTRIBUTIONS

---

I outline three principles for the specification of ideal point IRT models, making reference to important qualitative properties of the item response function as well as showing how the underlying principles affect estimation and interpretation of the model. Luo (1998, 2001) showed the most commonly used models to be part of a larger class of models he termed the “latitude of acceptance” models, in reference to the social judgment school (Sharif and Sharif, 1967). Luo showed that these models can be parameterized in terms of three quantities, all on the latent scale  $Q$ : object location  $b$ , subject location  $q$  and the latitude of acceptance  $a$ , which gives the magnitude of distance on the latent scale such that the probability of acceptance and rejection are equal. Choice of an operational function  $y(\cdot)$  determines the model. Heiser (1989) also set forth a general principle for the construction of ideal point models based on the transformation of a distance function,  $d(q,b)$ :  $P(q) = u \exp(-d(q,b)/a)$ , where  $u$  is the upper bound of the probability, which occurs when  $q = b$ . Heiser’s parameterization is very natural if one has a distance function in mind, but many useful models do not have particularly intuitive  $d(q,b)$ . A third principle involves rescaling the density of a random variable over  $Q$ . Because the usual normalization constraint for a pdf, it is necessary to renormalize the density to lie in the unit interval. This procedure has the substantial benefit of generating asymmetric ideal point models quite easily. For instance, the skew-normal density could be used (<http://azzalini.stat.unipd.it/SN/>) to form an item response function with asymmetric IRF. This approach also extends readily to a multidimensional  $Q$ .

**Vermunt, J.K.**

[SA-Monday-am-s3](#)

Tilburg University, Tilburg, The Netherlands ([j.k.vermunt@uvt.nl](mailto:j.k.vermunt@uvt.nl))

### Latent Class and Finite Mixture Models: Recent Developments

Though latent class and finite mixture models have a rather long history, only recently we see an increase in the application of these methods in psychological, sociological, educational, and biomedical research. Not only has the number of applied papers increased enormously, also the number of methodological contributions has grown tremendously the last ten to twenty year. Methodological contributions include papers on new models such as mixture regression models, mixture SEM and factor analysis, mixture growth models, latent class models with multiple latent variables, latent class behavioral and diagnostic models, and multilevel mixture models. Moreover, papers have been written on model selection issues (number of latent classes), improved algorithms to prevent local maxima, methods to check model identification, improved algorithms to speed up computations, etc. Another trend is the further integration between discrete (latent class models) and continuous (IRT, factor analysis, and random effects models) latent variable models. In this ‘state-of-the-art’ lecture, I will provide an overview of these recent developments.

**Verschoor, A.**<sup>(1)</sup>, Finkelman, M.<sup>(2)</sup>, Kim, W.<sup>(3)</sup>, and Roussos, L.<sup>(3)</sup>

[TS-Tuesday-pm-s48](#)

<sup>(1)</sup> Cito, Arnhem, The Netherlands ([Angela.verschoor@cito.nl](mailto:Angela.verschoor@cito.nl))

<sup>(2)</sup> Harvard School of Public Health, Boston, MA

<sup>(3)</sup> Measured Progress, Dover, NH

### Assembling Parallel Forms Alongside a Cognitive Diagnosis Model

Cognitive Diagnosis Models (CDMs) are designed to facilitate the measurement of examinee's strengths and weaknesses. In order to realize the full potential of CDMs, sound procedures for test assembly must be developed in the CDM framework. Several test assembly methods have been proposed for constructing a single test form alongside a CDM; these include the Cognitive Discrimination Index (CDI) and Genetic Algorithm (GA) approaches. However, the problem of assembling multiple parallel forms alongside a CDM remains open. In this paper, we introduce two new methods for assembling parallel forms alongside CDMs. Specifically, we extend the CDI and GA procedures to construct any prescribed number of forms. The problem of selecting an appropriate item overlap rate is explored. Multiple simulation sets compare the newly proposed methods to each other, as well as to forms selected at random. Practical uses of assembling parallel forms alongside CDMs are discussed.

**von Davier, M.**

IS-Wednesday-am-s60

Educational Testing Service, Princeton, NJ (mvondavier@ets.org)

Overview: Psychometric Modeling of Educational Survey Data

Large scale surveys of educational outcomes such as NAEP, TIMSS, PISA and IALS utilize data collection designs that are aimed at a broad coverage of a domain such as reading, mathematics, or science, while each respondent receives only a comparably small subset of the cognitive tasks representing the domain. The goal of survey assessments is not the reporting of individual scores, but rather the description of proficiency distributions in subgroups of interest, such as groups defined by gender, ethnicity and other variables. The collection of cognitive response data follows a complex matrix sampling design in order to realize the intended broad coverage, and combines the collection of responses to items in one of many booklets with a collection of background data in a context questionnaire, which asks the respondent about policy relevant variables. Models that integrate item response theory, with a population model that represents at least partially the complex populations assessed in the survey are commonly used to analyze data from these types of studies. The invited symposium presents studies that provide insight into innovative approaches to analyze large scale survey assessment data, either based on extensions of operational approaches, or based on latent variable models and models for classification of response patterns into knowledge spaces. These alternative approaches provide more in-depth understanding of the data by means of adding additional levels or more complex structures on the latent structure side as well as on the population model side, thus allowing differences between groups or clusters of respondents to help better understand these complex databases.

**von Davier, M., and Sinharay, S.**

CS-Tuesday-pm-s46

Educational testing Service, Princeton, NJ (MVonDavier@ets.org)

Stochastic Approximation Methods for Estimation of Latent Regression Item Response Models

This paper presents an application of a stochastic approximation EM-algorithm (SAEM) using a Metropolis-Hastings sampler to estimate the parameters of latent regression models. The utilization of stochastic approximation methods to obtain parameter estimates for high-dimensional models has been suggested by Robbins and Monro in 1951. Stochastic approximation is much less computationally involved than, for example, Monte-Carlo integration since it is based on a rationale that explicitly uses "noisy" evaluations of the objective function. Gu and Kong (1998) have demonstrated how to estimate mixed models using stochastic approximation. Deylon, Lavielle, and Moulines (1999) examined the convergence of SAEM, and more recently Cai (2007) suggested and applied stochastic approximation methods to item factor analysis. Latent variable regression models as used in large scale survey assessments are based on a multidimensional IRT model combined with a population model specified as a multidimensional, multiple regression of background variables on the latent variable. The methods put forward to estimate these high-dimensional models are computationally costly due to multidimensional integrals that need to be evaluated, or are based on technical approximations using assumptions that may be inaccurate for models that involve few item responses per dimension. Stochastic approximation methods using a Metropolis-Hastings sampler may serve as an alternative to those methods, since it is based on comparably small samples of draws from the actual posterior distributions that are involved. The applications of SAEM to estimating latent regression models presented in the talk are using large scale data from the National Assessment of Educational Progress (NAEP) from recent assessment cycles involving subject areas such as reading and mathematics, both of which are assessed with test booklets using multiple subscales.

**Walls, T.A.**, and Seong, J.  
University of Rhode Island, Kingston, RI (walls@uri.edu)

CS-Tuesday-am-s28

### Context Indices versus Time Indices in General Linear Mixed Models

Recent scholarship on the influence of context in psychological studies has focused on more explicit ways of modeling its influence (Little, 2007; Ram, 2007). In addition, consideration of the diversity of types of contexts, such as being in a setting with peers or relatives, a school or home setting, or a risky context has expanded in both theoretical and empirical literatures (Walls & Schafer, 2007; Walls, Jung & Schwartz, 2006). In modeling these intensive longitudinal studies, utilization of a time index is typical. However, the nature of the context may carry more meaning in relation to the processes under study than a time index enables. Moreover, such movements may take on variable meaning for one person versus another. Researchers may be able to assign scaleable values to the contexts, creating a time-varying index of contextual location. The development of such an index and its incorporation in the first level of a multilevel model is considered along with possible functional forms of the resulting model. The interpretation of covariates employed in higher levels of the model is considered. An example employing data from teenage smokers in a device-enabled diary study is described.

**Wang, A.**, and Cohen, A.S.  
University of Georgia, Athens, GA (wang0855@uga.edu)

CS-Monday-am-s8

### Evaluation of Three Test Speededness Models

The effects of unintended test speededness can potentially bias parameter estimates, possibly causing local dependence and threatening test reliability and even validity by introducing construct irrelevant variance. Different psychometric models have been developed to model test speededness, each of which provides a somewhat different interpretation of the effects. Three different speededness models which are useful for modeling speededness effects on paper-and-pencil tests are evaluated in this paper to provide researchers with some guidance about the effects of selecting one such model over others. In particular, the Hybrid model by Yamamoto, the two-class mixture model by Bolt et al. and the gradual change model by Goegebeur et al. are compared by applying each of the models to a large scale mathematical placement test. Classification consistency of examinees into the speeded and nonspeeded groups and analysis of patterns of responses of speeded and non-speeded examinees are examined across the three models.

**Wang, Chun**, and Chang, H.-H.  
University of Illinois at Urbana-Champaign, Champaign, IL  
(cwang49@cyrus.psych.uiuc.edu)

TS-Monday-am-s11

### Continuous $a$ -Stratification Index for Computerized Item Selection

The Continuous  $a$ -stratification Index (CAI) is proposed to improve the balance between estimation accuracy and item-exposure-control in the item selection process for large-scale Computer-Adaptive Tests (CAT).  $A$ -stratification, dividing the item pool into several strata and using lower- $a$  items in the early stages and higher- $a$  items later when the estimation is more accurate, has been one of the most successful item selection methods in the past decades. However, it might give rise to relatively large estimation error, because it arranges  $a$ -parameters in ascending order only between strata instead of item-by-item strictly. In contrast, the incorporation of CAI into item information could enable the item selection in a nearly strict  $a$ -ascending order. This is because CAI is maximized when the distance between the percentile of the  $a$ -parameter and the proportion of the exam completed is at a minimum. As a result, estimation accuracy and exposure balance will be improved; moreover, both of them can be optimized by adjusting the weight of CAI relative to item information. To conclude, CAI inherits the distinguished feature of the  $a$ -stratified method, and extends it to a continuous version. In our simulation study, CAI outperforms the  $a$ -stratified method in both estimation efficiency and item-exposure-control, thus the test reliability and security can be enhanced simultaneously.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Wang, Chunxin**, Ho, A., and Zhang, S.  
University of Iowa, Iowa City, IA (chunxin-wang@uiowa.edu)

Poster-Tuesday-pm-s50

### Modeling the Error Covariance Structure in Latent Growth Models

The purpose of this study is to explore the flexibility of Latent Growth Modeling (LGM) in modeling the error covariance structure. Various error covariance structures will be fit to a real longitudinal data set to better understand students' academic growth in mathematics and reading, as well as the factors influencing their growth. Four sets of latent growth models with different covariance structures will be fit into the data. Model 1 assumes that measurement errors are homoscedastic and independent within individuals over time. Model 2 relaxes the stringent assumption of homoscedasticity and allows measurement errors to be heteroscedastic. Model 3 assumes heteroscedastic measurement errors and that the errors were autocorrelated between certain time points. Model 4 allows an unstructured covariance model, where no constraints are imposed. This study demonstrates the flexibility of LGM in manipulating covariance structures in the context of modeling academic growth. More importantly, it offers a method for evaluating which covariates predict students' growth. As more and more states in the US start to pilot growth models in tracking students' academic development, these results may provide useful evidence to the educational practitioners and policy makers about the correlates of student growth.

**Wang, L.**, Pan, W., Bai, H., and Daniel, K.  
University of Cincinnati, Cincinnati, OH (Leigh.Wang@uc.edu)

IS-Tuesday-am-s33

### Power and Robustness of Multilevel Latent-Trait Differential Person Functioning: A Monte Carlo Comparison with Conventional Person Misfit Statistics

Research on aberrant response patterns that deviate from modeled prediction has received considerable attention in the psychometric literature. A recent line of inquiry uses the multilevel modeling approach for the detection of differential person functioning in latent trait models in a hierarchical data structure. Preliminary Monte Carlo studies comparing the performance of this innovative approach with conventional person misfit statistics have yielded encouraging results. However, systematic evidence is lacking in demonstrating its power performance and robustness across misfit conditions. The present study extends previous research by simulating three different types of person misfit: spuriously high, spuriously low, and random guessing, with gender bias as the group-level effect. The proposed approach along with three other conventional indices was then applied to the simulated data to compare their relative performance in recovering those misfitting simulees and detecting gender bias. The entire experiment was replicated 100 times to evaluate its stability performance. Preliminary results suggest that power performance of the proposed method is either superior or comparable to that of the conventional indices. Robustness performance, however, appears to be mixed among the four indices. Suggestions for future research to manipulate potential confounding or moderating factors are discussed.

**Wang, T.**<sup>(1)</sup>, and Hou, J.<sup>(2)</sup>  
<sup>(1)</sup> CASMA/University of Iowa, Iowa City, IA (tianyong-wang@uiowa.edu)  
<sup>(2)</sup> The School District of Palm Beach County, Florida, FL

IS-Tuesday-am-s32

### A Continuized Loglinear Approach to the Modified Frequency Estimation Equating Method under the Common-Item Non-equivalent Groups Design

Wang and Brennan (2008, APM) recently proposed a modified frequency estimation (MFE) equating method by modifying the basic assumption of the frequency estimation (FE) method. This MFE method has been shown to be effective in reducing the bias of the FE method while maintaining low standard error of equating (SEE). One difficulty is that when this method is combined with the traditional percentile rank based equipercentile equating, there is a need to find a conditional distribution conditioned on non-integer scores. The continuized log-linear (CLL) method which Wang (2008, APM) proposed as an alternative to the kernel continuization method (von Davier et al., 2004) can potentially overcome this difficulty by producing continuous bivariate distributions after the log-linear

## ABSTRACTS OF THE CONTRIBUTIONS

---

smoothing step. This paper propose to combine the MFE method with the CLL method under a common-item non-equivalent groups design and compare it to other equating method such as the kernel equating method under the same design.

**Wang, W.**, and Namgung, M.  
Wonkwang University, Jeollabuk do, South Korea (wj.weijie@gmail.com)

Poster-Tuesday-pm-s50

### Linking People's Perceptions and Physical Components of Sidewalk Environments via Rough Sets Theory

This paper provides a pilot study to develop a new approach on investigating people's perceptions on sidewalk environments by applying rough sets theory. A semantic-differential-technique based questionnaire survey was designed to collect 112 participants' psychological responses on 20 selected sidewalk environments in Iksan City of South Korea. A filed survey of the selected sidewalks was conducted to survey the physical components of the sidewalk environments. Through analysis participants' perception features were captured. Then the subjective and objective data obtained by two measures were combined together. Because conventional statistical methods are not appropriate due to the qualitative data, small sample size and uncertainty, the rough sets theory, an artificial intelligence technique, is applied to deal with the collected data. The application of the rough sets theory outputted the most important attributes to people's perceptions, minimal attribute sets without redundancy, and a series of decision rules that represented the relationships between perceptions and physical components of sidewalk environments. The analytical approach helps to better understand the people's perceptions to sidewalk environments in a small city and then establish a useful and constructive ground of discussion for walking environment design and management.

**Wang, Y.**, Ackerman, T., and Henson, R.  
University of North Carolina at Greensboro, Greensboro, NC  
(y\_wang2@uncg.edu)

Poster-Tuesday-pm-s50

### Factor Analytic Models and Cognitive Diagnostic Models: How Comparable Are They? - A Comparison of R-Rum and Compensatory MIRT Model with Respect to Cognitive Feedback

The necessity and importance of cognitive diagnosis is being realized by more and more researchers. There are a lot of models for cognitive diagnosis—the IRT-based discrete cognitive diagnosis models (ICDMs) and the traditional continuous latent trait models. However, there is a lack of literature that compares the newly defined ICDMs based on constrained latent class models to more traditional approaches such as multidimensional factor analysis. The purpose of this study is to compare factor models with the ICDMs with respect to cognitive feedback, thus providing useful feedback about model selection. Two models—R-RUM and 2PL CMIRT model—were selected for this purpose. Simulations were performed to explore the relationship between the estimated item parameters of the two models and to compare the feedback provided to examinees (examinee parameters). The results from the first study demonstrate that the two models are associated, but there is no one-to-one relationship between the two models in terms of test quality. Secondly, a method is proposed to compare the classifications from the ICDM to the continuous abilities estimated by factor analyses based on logistic regression. Subsequent comparisons were made within each model after estimating the R-RUM and the 2PL CMIRT, using common datasets.

**Wilderjans, T.**, Van Mechelen, I., and Ceulemans, E.  
K.U. Leuven, Belgium (tom.wilderjans@psy.kuleuven.be)

CS-Tuesday-pm-s46

### Simultaneous Analysis of Coupled Data Blocks that Are Subject to Different Amounts of Noise

In psychology, research questions often imply that different blocks of information pertaining to the same objects are to be analyzed simultaneously. As these blocks may emanate from different sources, often they may be subject to

## ABSTRACTS OF THE CONTRIBUTIONS

---

different amounts of measurement error or noise. To account for these differences in noise, different loss functions could be used in the data analysis. A challenging question then reads which loss function performs best with respect to the disclosure of the structure underlying the coupled data. To tackle this question, an extensive simulation study is carried out in which two types of loss functions are compared within the context of two models for coupled data that consist of two two-way two-mode data blocks with one mode in common: (1) a multiway multiblock components model for coupled real-valued data, and (2) a simultaneous clustering model for coupled binary data.

**Wilson, M.,** Zheng, X., and Walker, L.  
University of California, Berkeley, CA (markw@berkeley.edu)

IS-Tuesday-am-s34

### Latent Growth Item Response Models

In this presentation, we will review item response models that have been proposed for the analysis of longitudinal data, and relate that approach to the more standard one based on hierarchical linear modeling (HLM). One recent approach to combine these two approaches is to incorporate probit or logit links into HLM-type programs, thus allowing a direct representation of the item in the statistical model (e.g., see the work of Kamata and colleagues). However, there are limitations to the use of HLM-type programs, especially as it relates to issues concerning possible measurement complications, such as differential item functioning and multidimensionality. We describe an alternate approach that formulates the model somewhat differently, and allows one to take advantage of the strengths of generalized item response model software such as *ConQuest*, SAS NLMIXED and *gllamm*. This elaborates on the work of Glas incorporating time-based DIF into measurement models. We demonstrate this approach with an analysis of NELS88 data, including a simulation based on the NELS context. We conclude the presentation by describing some of the interesting hypotheses that can be addressed using this approach.

**Willson, V.L.,** Kwok, O.-M., and Liew, J.  
Texas A&M University, College Station, TX (v-willson@tamu.edu)

CS-Monday-pm-s19

### Construct Noninvariance in Growth Modeling

The noninvariance of constructs over time is examined in estimating growth in subjects. We propose a general framework for estimating growth based on rescaling of variances over time. The models we consider include: 1) Single measure for a construct with change in measurement; 2) Two measures with change in measurement of one measure; 3) One latent factor with change in measurement; and 4) two latent factors with change in measurement in one factor. We assume in each case at least 3 measurements in an initial measurement invariant time-frame, followed by at least two measurements in a second, different measurement-invariant time-frame. We propose a two-step procedure in which the measurement-invariant properties are established for each time frame. A preliminary growth model is fit for the initial time-frame, and a predicted value estimated for the next time point past the initial time-frame. The factor structure of the second-time frame is corrected by equating factor variance to be equal to the pooled variance of the initial time-frame factor. A spline fit is made to the entire model. Covariate processes are added as predictors with either time-invariant or growth processes themselves.

**Wise, S.L.,** and DeMars, C.E.  
James Madison University, Harrisonburg, VA (wisel@jmu.edu)

IS-Monday-pm-s18

### An Investigation of Rolling Person Fit in Identifying Examinees Who Abandon Test Effort

There has been an increased interest in examinee test-taking motivation, especially in situations where there are no personal consequences for test performance. It has been shown, using item response time, that some examinees exhibit good effort during the early stages of a test, but at some point these examinees will begin to exhibit non-effortful behavior. This study examined the utility of person fit indices in identifying instances when examinees have changed response strategies from effort to non-effort. It was found that a rolling likelihood ratio statistic can,

## ABSTRACTS OF THE CONTRIBUTIONS

---

under some conditions, effectively identify response strategy changes. Implications of these findings for improving test score validity are discussed.

**Woods, C.**

IS-Wednesday-am-s53

Washington University, St. Louis, MO (cwoods@artsci.wustl.edu)

### Ramsay-Curve Differential Item Functioning

In Ramsay-curve item response theory (RC-IRT; Woods & Thissen, 2006), logistic item response functions are fitted to data using marginal-maximum likelihood estimation simultaneously with estimation of the latent density using splines. Previously, it has been possible to use RC-IRT only for a single sample of people. The present research describes how RC-IRT can be used for IRT-based likelihood-ratio tests of differential item functioning (IRT-LR-DIF). An implementation of IRT-LR-DIF is described wherein the latent density is presumed standard normal for the reference group but approximated as a Ramsay curve for the focal group. The new procedure, RC-DIF-1, is designed for types of variables that are approximately normal for the majority or mainstream group with which the test was constructed, but possibly not for some focal groups. This could occur when the focal group is more heterogeneous than the reference group. For example, English reading proficiency may be approximately normal for native English speakers but not for nonnative English speakers, which may include persons who speak many different languages and have many different cultural backgrounds. Simulation results will be presented showing that when the focal group density is nonnormal, RC-DIF-1 provides more accurate results than standard IRT-LR-DIF (for which both densities are presumed normal).

**Wu, H.,** and Browne, M.W.

TS-Tuesday-am-s31

The Ohio State University, Columbus, OH (wu.498@osu.edu)

### An Empirical Bayesian Approach to Misspecified Covariance Structures

“All models are wrong, but some are useful.” (Box, 1976). Because models typically do not fit exactly to the population in real life applications, the classical null hypothesis testing procedure suffers from the fact that the null hypothesis will eventually be rejected given a large enough sample. To avoid this problem, various measures of model misfit have been developed. These measures of misfit, however, are Post-hoc modifications of the likelihood ratio test statistic for a perfectly fitting model. To avoid this self-contradictory *modus operandi*, I shall present an empirical Bayesian approach, developed in conjunction with Dr. Michael W. Browne, which directly addresses the issue that the model does not fit the population covariance matrix. In this approach, we model two different aspects of discrepancy between the observed sample and the structured model. In addition to the sampling errors which result in the difference between the sample and the population, systematic errors which give rise to the discrepancy between the population and the model are modelled by a prior distribution on the population covariance matrix. An additional parameter denoting the dispersion of the prior distribution is considered as a measure of misspecification and estimated together with the covariance structure.

**Wu, J.-Y.,** Kwok, O.-M., and Hsu, H.-Y.

CS-Wednesday-am-s57

Texas A&M University, College Station, TX (jyunyu@neo.tamu.edu)

### Comparing the Efficiency of Robust Estimators and Multilevel Models on Analyzing Multilevel Data - A Monte Carlo Study

Multilevel data with non-independent observations are very common in social sciences. Ignoring the dependency issue results in bias estimation of the standard error of the fixed effect (or regression coefficient), which in turn, affects the statistical inference of the fixed effect. One way to handle the non-independent observations in structural equation models (SEMs) is to adopt the ad-hoc robust sandwich standard error estimators, such as Huber-White corrected standard error estimator (e.g., the Type=Complex routine in Mplus). Another way to handle the non-

## ABSTRACTS OF THE CONTRIBUTIONS

---

independency issue is to analyze the data using multilevel models (e.g., the Type=Twolevel routine in Mplus). In this paper, we examined the similarity and difference on analyzing data with non-independent observations between these two approaches—the Type=Complex routine assumes equal between and within models while the Type=Twolevel allows different between and within models. A Monte Carlo study was conducted to evaluate the above scenarios, with considering factors including sample size, cluster size, intra-class correlation, and degrees of model misspecification. Implications of the findings are discussed.

**Wyse, A.E.**

TS-Tuesday-am-s38

Michigan State University, East Lansing, MI (wyseadam@msu.edu)

### IRT Theory for DIF Cancellation

A recent development in the study of differential item functioning (DIF) is the possibility of DIF cancellation on a full-length assessment. With this goal in mind, statistical procedures have been developed that can empirically test for DIF cancellation. However, the underlying theory of when DIF cancellation is possible is not fully understood. This article develops psychometric theory for DIF cancellation and discusses the conditions under which DIF cancellation can occur using an IRT framework. It is shown that complete DIF cancellation is only possible when each item given to the focal group can be matched to an item given to the reference group that has identical item parameters. The theoretical implications of these results are discussed and examples of the potential impact of the lack of DIF cancellation are illustrated as they relate to student proficiency classifications.

**Xu, S., and Blozis, S.A.**

CS-Tuesday-am-s30

Department of Psychology, University of California, Davis, CA  
(shuxu@ucdavis.edu)

### A Two-Part Mixed-Effects Model for Time-Use Data: A Comparison of Maximum Likelihood and Bayesian Estimation

Longitudinal time-use data often present analytic challenges due to the high frequencies of zero. A two-part mixed-effects model may be applied to address the semi-continuous distribution of these and similar behaviors (Olsen & Schafer, 2001). The model is based on a joint distribution of a continuous and a categorical response. Based on repeated measures of reading and TV use for a large sample of children, a two-part mixed-effects model is applied to study the joint associations between the behaviors. This study compares maximum likelihood and Bayesian estimation procedures. Results are compared in terms of parameter estimates, convergence rates, and computing time.

**Xu, X., and Jia, Y.**

IS-Wednesday-am-s53

Educational Testing Service, Princeton, NJ (xxu@ets.org)

### On Sensitivity of the Latent Ability Distribution

Abstract: Latent ability distribution is a fundamental concept in item response theory (IRT), and plays an important role in the field of educational testing. A normal distribution is usually assumed for the latent ability distribution in practice partly because (1) it is simple and well understood; (2) it is fully determined by its first two moments. However, there are situations where a normal distribution assumption might not be inappropriate (Dresher & Moran, 2002). In this study, we propose to apply generalized skewed normal (GSN) distribution to describe the latent ability. The GSN is a more flexible distribution which includes normal distribution as a special case. After introducing the GSN distribution, a simulation study is carried out. Rasch and two parameter logistic (2-PL) IRT models coupled with various shapes of latent ability distribution are applied to generate the item responses. In addition, simultaneous estimation of both item parameters and distributional parameters for the latent ability are employed. The effects of misspecification of the latent ability distribution in the estimation procedure are examined

## ABSTRACTS OF THE CONTRIBUTIONS

---

in terms of the item parameter estimates as well as the distributional parameter estimates. The results show that latent ability distribution misspecification affects the quantile recovery of latent ability distribution, but does not considerably affect the item parameter estimates.

**Xu, X.**, and von Davier, M.  
Educational Testing Service, Princeton, NJ (xxu@ets.org)

IS-Wednesday-am-s60

### NAEP Reading Data Analysis Using GDM Framework

The National Assessment of Educational Progress (NAEP) has long been regarded as the gold standard to measure the academic progress for U.S. students in grades 4, 8 and 12. A partially balanced-incomplete-block (pBIB) design is employed to ensure that the assessments cover a wide range of content in subjects of interest. In the NAEP reading assessment, this design leads to the fact that individual students take only 15-30 items for two or three subscales. This small amount of items obviously cannot lead to an accurate estimation of individual abilities. Various models (Mislevy, 1992; Aitkin & Aitkin, 2004; Xu & von Davier, 2006; von Davier, 2007; de la Torre, & Douglas, 2006; Li & Oranje, 2006) have been suggested to compensate the shortage of the cognitive items. All these models attempt to impose constraints or assume structures on the space spanned by the latent ability variable(s). The primary goal of this presentation is to compare different latent variable models, as specified in terms of latent structures of differing dimensionality, as well as different assumptions about the population structure governing the ability space, using the general diagnostic model (von Davier, 2005) framework. All analyses presented are based on real data from recent NAEP reading assessments, comprising of about 150,000 to 180,000 students, each assessed with one of about 20 subsets (booklets), with the whole set covering a total of about 300 reading tasks. Through this analysis, we hope to demonstrate the advantages and disadvantages of different models as well as give some recommendation for future applications. In the mean time, the large scale assessment programs may benefit from sharing our experience with the NAEP data.

**Yamamoto, K.**  
Educational Testing Service, Princeton, NJ (kyamamoto@ets.org)

IS-Wednesday-am-s60

Discussion of the session “Advances in Psychometric Modeling of Large Scale Educational Survey”

**Yang, M.**, and Chow, S.-M.  
University of Notre Dame, Notre Dame, IN (myang@nd.edu)

CS-Wednesday-am-s63

### Using State-Space Models with Regime Switching to Represent the Dynamics of Facial EMG Data

Facial electromyography (EMG) is a useful physiological measure for detecting subtle affective changes in real time. It can differentiate valence of emotion, capture transient and covert affective response, and obtain an approximately continuous-time measure of emotion change. In this study, facial EMG data is analyzed using time series analysis methods. By allowing certain parameters to switch between several discrete stages (regimes), regime-switching models can be used to describe various transition patterns, such as sudden shifts, gradual changes and heteroskedastic variances. The main purpose of this study is to construct and propose different regime-switching state-space models suited for representing the time-varying dynamics of facial EMG data an relationship with emotion regulation process. The Kim filter, which is an extension of the Kalma filter, is proposed to estimate the latent states and the Gaussian maximum likelihood method is use for parameter estimation. The change patterns of EMG data from different experimental condition will be analyzed and compared.

## **ABSTRACTS OF THE CONTRIBUTIONS**

---

**Yang, N.,** and Habing, B.  
University of South Carolina, Columbia, SC (genuion@hotmail.com)

TS-Wednesday-am-s59

### Identifying Item Type on Mixed Unfolding/Monotone Instruments

Educational assessments are primarily composed of monotone (or dominance) items and are commonly analyzed using the Rasch, 3PL or partial credit model. Instruments in other fields may, however, be composed of unfolding (or proximity) items which can be analyzed using models such as the GGUM. At present, the procedures for fitting these two item types can only be applied when the exam is composed entirely of one of the item types. This talk is a first step towards developing a procedure to simultaneously estimate both classes of models for an exam where both item types are present. In particular, it examines the ability to classify the items on a mixed exam into the appropriate type using methods based in classical test and scaling theory. It also examines the performance of those methods in determining initial model starting values for use in subsequent maximum likelihood estimation.

**Yang-Wallentin, F.,** Jöreskog, K.G., and Luo, H.  
Norwegian School of Management, Uppsala University, Sweden  
(fan.yang@dis.uu.se)

TS-Tuesday-am-s31

### Confirmatory Factor Analysis with Categorical Data: An Evaluation of Different Estimation Methods

Categorical measures are common in many empirical investigations in the social and behavioral sciences. Researchers often apply Maximum Likelihood Confirmatory Factor Analysis (CFA) which assumes that these measures have normal distributions. A better approach is to use polychoric correlations and fit the models using some robust method such as Robust Unweighted Least Squares (RULS), Robust Maximum Likelihood (RML), Weighted Least Squares (WLS), or Diagonally Weighted Least Squares (DWLS). In this simulation evaluation we study the behavior of RULS, RML, WLS, and DWLS in combination with polychoric correlations with both normal and non-normal underlying variables.

**Yao, L.,** and Rich, C.  
CTB/McGraw-Hill, Monterey, CA (Lihua\_Yao@ctb.com)

TS-Wednesday-am-s59

### Application of TestLet Effect Model to Performance Assessments with Multiple Scoring Rubric

Testlet effect in scaling has been investigated by many researchers when several items that are related to a common stimulus. Because the information used to answer these items in the passages is often interrelated, the traditional IRT model has been modified to account for the testlet effects. A recently published English language proficiency speaking test is scored using multiple-criteria scoring rubric in which one response is scored two or three times using different criteria such as grammar and meaning. This type of testlet is different from testlet resulted from items with a common stimulus. It is similar to trait analytical scoring found in writing assessments where a writing response is scored by several traits. Testlet based tests and performance assessments are known to be likely to produce local item dependence, a violation of item local independence assumption in IRT scaling. In this study, multidimensional two parameter partial credit models and two parameter partial credit testlet-effect models are proposed to investigate the testlet effects found in performance assessments. The goal of this research is to provide improvement in accuracy of parameter estimation and proficiency estimation for multiple-criteria scored test data through both real data and simulation data.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Ye, F.**<sup>(1)</sup>, and You, W.<sup>(2)</sup>

CS-Wednesday-am-s57

<sup>(1)</sup> University of Pittsburgh, Pittsburgh, PA (feifeiye@pitt.edu)

<sup>(2)</sup> Pearson Education, Upper Saddle River, NJ

### Multilevel IRT Model and Multilevel SEM in Estimating the Effect of Multilevel Covariates on a Latent Trait Measured by Dichotomous Items

The effect of multilevel covariates on a latent trait which is measured using dichotomous or polytomous items can be examined using three different approaches: 1) multilevel IRT with items at level 1, students at level 2, and contextual information (e.g., school variables such as SES, sector size) at level 3; 2) multilevel SEM, and 3) a two-step procedure which first uses IRT model to estimate student latent trait score while ignoring the hierarchical data structure and then fit a two-level hierarchical linear model on the latent scores. This study addresses dichotomous items only and aims to compare these three approaches on parameter recovery of fixed effects and random effects using a Monte Carlo simulation study with varied number of items, student-level and school-level sample sizes, intra-class correlations, and effect sizes. The IRT models considered are 1-parameter and 2-parameter logistic models.

**Yu, H.-T.**, and Vermunt, J.K.

CS-Wednesday-am-s63

Leiden University, Leiden, The Netherlands (hsiutingyu@gmail.com)

### Interpret and Test the Latent Classes Transition of the Mixed Latent Markov Models

The transition matrix in the Mixed Latent Markov Models provides a probabilistic description of transitions between discrete latent classes. The interpretation of transition between latent classes can become very complex when there are more than two latent classes. The conventional approach of describing the transition is to set up a reference class as a baseline; however, it is not intuitive under the special nature of transition matrix. In the exploratory aspect, we explore and discuss different approaches of setting up a reference class for the interpretation of transition among different latent classes. We show that some special reference schemes can be set up to facilitate the description and interpretation of the latent transition process. In the confirmative aspect, we discuss how to set up and test some specific hypotheses regarding the transition pattern. This confirmative aspect provides a valuable tool for the empirical application. Both exploratory and confirmatory approaches of describing and testing transition between latent classes will be demonstrated using an empirical example. We will also present how various approaches can be set up using the Latent Gold 4.5 syntax module.

**Yung, Y.-F.**

CS-Tuesday-pm-s47

SAS Institute Inc., Cary, NC (Yiu-Fai.Yung@sas.com)

### Testing and Contrasting Mediation or Indirect Effects in SEM: An Analytic Approach and its Implementation

Testing and contrasting mediation or indirect effects offers researchers more insights about the causal processes of their structural equation models. In this talk, we summarize an analytic approach, which is based on the Wald-type tests, to the problem. We will show a computationally efficient method that can evaluate the standard error estimates of indirect and total effects "on demand." Hence, evaluating large sparse matrices in Kronecker products is avoided in the implementation. To illustrate the method, we will show the results of an example that is analyzed by the TCALIS procedure in SAS/STAT 9.2.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Zhang, G.**<sup>(1)</sup>, Ong, A.D.<sup>(2)</sup>, and Bergeman, C.S.<sup>(1)</sup>

IS-Wednesday-am-s52

<sup>(1)</sup> University of Notre Dame, Notre Dame, IN (gzhang3@nd.edu)

<sup>(2)</sup> Cornell University, Ithaca, NY

### A Sandwich-Type Standard Error Estimator for Vector Autoregressive Moving Average Models

Multivariate time series provides opportunities for studying causal relations. The vector autoregressive moving average (VARMA) model is a popular choice for multivariate time series. In Psychology, the VARMA model has been used in repeated time series and single time series, manifest variable time series and latent variable time series. Fitting the VARMA model to lagged correlation matrices obtained from a single subject presents a special challenge because these matrices do not have a Wishart distribution. In particular, the standard errors and test statistic may be invalid in this situation. In the talk, we propose a sandwich type standard error estimator. Another benefit of the sandwich-type estimator is that the model can be 'imperfect' in the population. The procedure will be illustrated using a study on emotion change patterns in older adults.

**Zhang, J.**

CS-Tuesday-am-s36

Educational Testing Service, Princeton, NJ (jzhang@ets.org)

### Asymptotic Normality of DETECT Index and a New Significance Test for Unidimensionality

DETECT, short for *dimensionality evaluation to enumerate contributing traits*, is a statistical procedure that is used to identify the number of dominant latent dimensions and to estimate the degree of multidimensionality. In this study, the asymptotic normality of DETECT statistics was proved and its asymptotic standard error (SE) was formulated. Based on the asymptotic theory, standardized DETECT statistics were constructed, which can be used to test statistical hypothesis,  $H_0$ : Unidimensional vs.  $H_1$ : Multidimensional, for response data. The pre-asymptotic performance of the statistic was optimized by making an adjustment on the asymptotic SE based on simulation studies. The statistical test was then applied to simulated response data with different dimensional structures and with various numbers of items and examinees. The simulation study indicates that this significant test has a very small rate of type II error while controlling the rate of type I error under a pre-selected significant level (e.g., 0.05). The results were also compared with those obtained using the original DETECT procedure. This comparison shows that the new standardized DETECT dramatically reduced the error rates in identifying unidimensionality versus multidimensionality.

**Zhang, S.**, and Nozawa, Y.

Poster-Tuesday-pm-s50

University of Iowa, Iowa City, IA (su-zhang-1@uiowa.edu)

### Applying Bootstrap and Jackknife to Estimate Variability of Variance Components in Generalizability Theory

In generalizability theory (Brennan, 2001) accurate and precise estimation of variance components is crucial. The resampling procedures, such as the bootstrap and the jackknife (Efron, 1982), have been employed to assess the bias and variability of such statistics in G-theory (Tong & Brennan, 2006, 2007). It has been found that the jackknife procedure performed better than the bootstrap and that the bootstrap estimates of variance components were seriously biased. The purpose of this study is two-fold: first, to apply both the bootstrap and the jackknife procedures to assess the bias and the variability of estimated variance components; second, to investigate Brennan's (2007) bias-correction formulas for bootstrap estimates and to explore how the corrected/adjusted bootstrap estimates are affected by the specific bootstrap sampling plan employed, the pattern of sample sizes, and the pattern of variance parameters. A series of simulation studies were conducted to obtain the unadjusted bootstrap variance estimates (i.e., without applying the bias correction formulas by Brennan (2007)), the adjusted bootstrap variance estimates, the jackknife variance estimates, and the standard errors associated with each procedure. The single-facet crossed design ( $p \times i$ ) was the focus in this study, with score effects assumed to be multivariate normal.

## ABSTRACTS OF THE CONTRIBUTIONS

---

**Zhang, T.**, and Rupp, A.A.  
University of Maryland, College Park, MD (zhangt@umd.edu)

TS-Tuesday-pm-s48

Sensitivity of Parameter Recovery and Classification Accuracy for Cognitive Diagnosis Models to Prior Specification under a Bayesian Estimation Framework

Despite an increasing number of Bayesian estimation algorithms for *cognitive diagnosis models (CDMs)* (e.g., diBello, Roussos, & Stout, 2007), few studies have investigated the sensitivity of CDMs to different priors for various estimation conditions. In this study, using R and Winbugs, we investigate how different types of priors (conjugate vs. non-conjugate) and different parameter specifications for these priors (informative vs. non-informative) influence the recovery of item parameters and classification accuracy for the DINA and DINO models (e.g., Rupp & Templin, 2007) under varying sample sizes (small vs. moderate), number of attributes (few vs. many) and loading structures (simple vs. complex). Results from this study will provide insight into the degree to which a theory-driven specification of priors can meaningfully improve upon model estimation.

**Zhang, Y.-P.**  
National Pingtung University of Education, Taiwan (peng@mail.npue.edu.tw)

Poster-Tuesday-pm-s50

TAT Pre-Referral Intervention for Students with Learning Difficulties & Behavioral Problems

The purpose of this study is to seek the problems and to understand the processes of solving problems for interviewing 16 general teachers and 26 students with learning difficulties and behavioral problems in the general classrooms located in five different schools. In addition to investigate the effectiveness and influenced factors of the TAT pre-referral intervention for those students by the action research. The main findings are the majority of 26 students made a progress in their attention, writing ability (50%), emotional (31%) and learning attitude (42%). Moreover, some of 26 students were suspected having learning disabilities (62%), language disorders (23%). They need individual assistance (42%), peer-tutoring in learning mathematics and language arts (23%) as a remedial instruction. Furthermore, the majority of the students made positive progresses (89%), particularly in language arts and mathematics within one year in the pre-referral intervention; but seldom students had negative progresses.

**Zhang, Z.**<sup>(1)</sup>, Takane, Y.<sup>(1)</sup>, and Lu, J.<sup>(2)</sup>  
<sup>(1)</sup> McGill University (zhidong.zhang@mail.mcgill.ca)  
<sup>(2)</sup> Univ Hong Kong

Poster-Tuesday-pm-s50

Modeling Situation Interpretations and Cognitive Trajectories in a Clinical Problem Solving Process with Bayesian Network.

Modeling cognitive processes in clinical learning environments is a necessary first step toward improving assessment in this domain. Verbal protocol and cognitive content analyses are effective methods of exploring such cognitive processes, and for the purpose of simplifying the discussion, we have labeled these processes as Recognizing Information, Deep Cognition, and Cognitive Action. Exploring clinical problem solving processes with Bayesian network techniques can characterize students' dynamic learning processes quantitatively, identify differences in cognitive components at different stages of learning and better represent clinical problem solving features. We describe here a hierarchical cognitive model that can be used to describe the complex cognitive network relations among various clinical cognitive components. This model can then be used to dynamically update students' moment-by-moment progress through a clinical case. The study concludes that the cognitive model was useful in identifying students' learning trajectories by representing the different cognitive features chronologically.

**Zijlstra, W.P.**, van der Ark, L.A., and Sijtsma, K.  
Tilburg University, Tilburg, The Netherlands (w.p.zijlstra@uvt.nl)

CS-Wednesday-am-s62

### Detecting Various Types of Outlying Behavior in Questionnaire Data

Several methods are available for outlier detection in multi-item questionnaire data used for measuring psychological constructs such as introversion and neuroticism. In this paper, we investigated four outlier detection methods that are based on few model assumptions. The methods were the Mahalanobis distance, a nonparametric distance-based approach, and two methods proposed by Zijlstra et al. (2007). Simulation studies were performed in which the data consisted of item-score vectors from regular and outlying simulees. Outlying item-score patterns were generated by either different response styles, different person response functions, or different latent trait distributions. The specificity and the sensitivity of the outlier detection methods were determined. Furthermore, the influence of the outliers on the norm distribution of the questionnaire scores and on psychometric statistics (e.g., Cronbach's  $\alpha$ ) was investigated.

**Zu, J.**, and Chow, S.-M.  
University of Notre Dame, Notre Dame, IN (jzu@nd.edu)

CS-Tuesday-am-s28

### Examining Dynamic Factor Models with Time-Varying Parameters

To better investigate complex nonstationary multivariate time series in psychology, the current project aims to extend earlier work on dynamic factor models to cases involving time-varying parameters. A method is proposed that uses the extended Kalman smoother (EKS) to estimate factor scores and track time-varying parameters, the Gaussian maximum likelihood (GML) for point estimation of other time-invariant parameters and the bootstrap for standard error (SE) estimation. The performance of the proposed method and the consequences of violating the stationary assumption were examined via Monte Carlo simulations, in the context of a dynamic one-factor model with one nonlinear time-varying autoregressive coefficient. Results suggest that (i) the EKS recovered dynamics of latent factors and time-varying coefficients rather faithfully; (ii) the GML estimator, standard asymptotic and bootstrap SEs worked well for most parameters. However, they were biased for certain parameters in the dynamic model; (iii) the amplitude of changes of the time-varying coefficient played a crucial role in determining whether the coefficient can be assumed to be fixed and the accuracy of estimates of parameters and SEs. Explanations and practical implications for studies of human dynamics are provided.

INDEX

- Abad, F.J., 2  
 Abdi, H., 75  
 Ackerman, T., 4, 88  
 Adachi, K., 1  
 Ahmed, U.S., 1  
 Aitchison, K.J., 79  
 Allaire, J.C., 11  
 Anderson, C., 46  
 Asparouhov, T., 56  
 Bahn, G., 1  
 Bai, H., 87  
 Barrada, J.R., 2  
 Bartolucci, F., 60  
 Bauldry, S., 3  
 Béguin, A., 2  
 Bentler, P., 46  
 Bergeman, C.S., 95  
 Blais, J.-G., 62  
 Blozis, S.A., 2, 11, 91  
 Bohn, C.M., 17  
 Bollen, K.A., 3  
 Bolt, D.M., 3  
 Boomsma, A., 76  
 Borsboom, D., 3  
 Bouwmeester, S., 3  
 Bovaird, J.A., 4  
 Braeken, J., 4  
 Brasfield, J., 4  
 Brosschot, J.F., 81  
 Brossman, B., 5  
 Browne, M.W., 5, 90  
 Bryant, F., 1  
 Budescu, D.V., 34  
 Burns, T., 63  
 Büyükkurt, B.K., 40  
 Cai, L., 6, 77  
 Cardinale, J., 6  
 Carroll, J.D., 19  
 Carstensen, C.H., 6  
 Casabianca, J., 7  
 Cella, D., 7  
 Ceulemans, E., 88  
 Chajewski, M., 7  
 Chang, C.-H., 47  
 Chang, H.-H., 1, 47, 86  
 Chang, S.-W., 8  
 Chen, P.H., 8  
 Chen, Q., 8, 9, 50  
 Cheng, Y., 9  
 Chiu, C.Y., 9  
 Cho, S.-J., 10  
 Choi, H.-J., 10  
 Choi, J., 24  
 Choi, S.W., 75  
 Choulakian, V., 10  
 Chow, S.-M., 11, 92, 97  
 Chumney, F., 4  
 Clavel, J.G., 11, 57  
 Coffman, D.L., 11  
 Cohen, A.S., 45, 86  
 Conijn, J.M., 12  
 Corter, J.E., 13  
 Croudace, T., 12, 63, 65  
 Dai, H., 8  
 Daniel, K., 87  
 Daniel, R.C., 13  
 De Boeck, P., 4, 13, 49  
 De Champlain, A., 32  
 de la Torre, J., 14, 27  
 de Rooij, M., 14, 60  
 Dean, M.J., 13  
 DeMars, C.E., 89  
 Ding, S., 8, 14, 50  
 Dogan, E., 15  
 Dolan, C.V., 33, 54  
 Dorans, N., 15  
 Douglas, J., 9  
 Dumenci, L., 15  
 Dunn, J., 16  
 Edwards, M.C., 16  
 Embretson, S.E., 16, 52  
 Emons, W.H.M., 12, 17  
 Erosheva, E.A., 17  
 Fargo, J.D., 17  
 Feldman, B.J., 18  
 Ferdous, A., 65  
 Finch, H., 18  
 Finkelman, M., 16, 18, 38, 84  
 Foss, T., 58  
 Fox, J.P., 39

## ABSTRACTS OF THE CONTRIBUTIONS

---

- Fox, J.-P., 19  
France, S.L., 19  
Frederickx, S., 49  
Frees, E.W., 38  
French, B., 18  
Fukunaka, K., 20  
Furgol, K., 26  
Garcia, R., 20  
Geerlings, H., 20  
GENDEP Consortium, 79  
Gierl, M.J., 44  
Glas, C.A.W., 20, 37  
Golding, J., 65  
Golombok, S., 65  
Gonzalez, E.J., 21  
Goodrich, B., 21  
Grasman, R.P.P.P., 23  
Grelle, D., 21  
Griffiths, T., 22  
Groenen, P.J.F., 83  
Gueorguieva, R., 22  
Gundula, A.M., 22  
Habing, B., 72, 93  
Hafdahl, A.R., 23  
Halpin, P.F., 23  
Hamaker, E.L., 11, 23  
Han, K.T., 24, 42  
Hancock, G.R., 24  
Hansson, L., 63  
Harring, J.R., 24  
Harris, D., 25  
Hashimoto, T., 59  
Hauser, C., 39  
Hayashi, K., 25  
Heiser, W.J., 60  
Henson, R., 88  
Heo, M., 25  
Herrick, R.C., 75  
Hessen, D.J., 26  
Hines, M., 65  
Ho, A., 26, 87  
Hofmans, J., 26  
Holland, P.W., 27, 71  
Holling, H., 20  
Hong, Y., 27  
Horton, D., 28  
Hoshino, T., 53, 58  
Hou, J., 87  
Hoyer, P.O., 70  
Hsieh, J.-C., 38  
Hsu, H.-Y., 28, 90  
Hwang, H., 28, 31, 74, 75  
Hyvarinen, A., 70  
Ikehara, K., 29  
Iliopoulos, G., 34  
Ip, E., 29  
Ishioka, T., 59  
Iwama, N., 29  
Jahng, S., 30  
Jansen, M.G.H., 30  
Javaras, K.N., 30  
Jeltova, I., 6  
Jia, Y., 91  
Joe, H., 51  
Johnson, T.R., 3  
Jöreskog, K., 31  
Jöreskog, K.G., 58, 93  
Jung, K., 31  
Jung, Song, 31  
Jung, Sunho, 32  
Junker, B.W., 32, 83  
Kahraman, N., 32  
Kallert, T., 63  
Kan, K.J., 33  
Kano, Y., 25, 40, 75, 79  
Kaplan, D., 33  
Karelitz, T., 34  
Kateri, M., 34  
Kato, K., 34  
Kawahashi, I., 35  
Kelderman, H., 35  
Keller, L.A., 16, 35, 36  
Keller, R., 16, 35, 36, 60  
Kelly, K., 36  
Kenny, D.A., 20, 36  
Khalid, M.N., 37  
Kim, I.-H., 37  
Kim, J.S., 38  
Kim, K.H., 38  
Kim, S., 43  
Kim, W., 16, 18, 38, 84  
Kim, Y. Y., 15, 39  
Kingsbury, G.G., 39  
Klein Entink, R., 39

## ABSTRACTS OF THE CONTRIBUTIONS

---

- Klinkenberg, S., 82  
Koning, A.J., 83  
Konya, Y., 40  
Kreager, D., 17  
Kroonenberg, P.M., 40  
Kroopnick, M.H., 24  
Kubo, S., 29  
Kuha, J., 40  
Kunina, O., 41  
Kuppens, P., 26  
Kupzyk, K.A., 41  
Kwok, O.-M., 9, 28, 43, 89, 90  
Kyungtae, K., 41  
Lamis, D.A., 50  
Lane, S.P., 42  
Lawrence, D.R., 42  
Ledgerwood, A., 70  
Lee, J., 42  
Lee, M. D., 43  
Lee, S., 43, 53  
Lee, W.-C., 5, 43, 44  
Lee, Y.-H., 43, 44  
Lei, M., 44  
Leighton, J.P., 44  
Lewis, C., 7  
Li, D., 45  
Li, F., 45  
Li, L., 46  
Li, Z., 46  
Liang, L., 5  
Liew, J., 89  
Ligtvoet, R., 46  
Lin, A., 46  
Lin, H., 47  
Lin, N., 47  
Linting, M., 48  
Liu, J., 48  
Long, D., 73  
Loye, N., 48  
Lu, Z., 49  
Luo, F., 14  
Luo, H., 93  
Luo, W., 9  
MacKinnon, D.P., 78  
Magda, T., 26  
Magidson, J., 49  
Magis, D., 49, 62  
Malone, P.S., 50  
Mao, M.-M., 50  
Maraun, M.D., 23  
Maris, G., 50, 82  
Markus, K., 28  
Marsiske, M., 11  
Masyn, K.E., 18, 50, 51  
Matsueda, R.L., 17  
Mavridis, D., 51  
Maydeu-Olivares, A., 51  
Mayekawa, S.I., 52, 58  
McGuffin, P., 79  
McIntyre, H.H., 52  
Merkle, E.C., 52  
Mesman, J., 40  
Millsap, R.E., 53  
Miyazaki, K., 53  
Miyazaki, Y., 53  
Molenaar, D., 54  
Mooijaart, A., 54  
Moses, T.P., 15  
Moustaki, I., 54  
Murakami, T., 55, 56  
Muthén, B., 56  
Nakamura, K., 56  
Namgung, M., 88  
Nering, M., 16, 38  
Nishisato, S., 11, 57  
Noel, Y., 57  
Northrup, T., 50  
Nozawa, Y., 95  
Ntzoufras, I., 34  
Oh, H., 57  
Okada, K., 58, 69  
Okubo, T., 58  
Olea, J., 2  
Olsson, U.H., 58  
Ong, A.D., 95  
Oshima-Takane, Y., 76  
Otsu, T., 59  
Paccagnella, O., 59  
Pan, W., 87  
Papanastasiou, E.C., 59  
Park, S.-H., 73  
Parker, P., 60  
Pennoni, F., 60  
Perroud, N., 79

## ABSTRACTS OF THE CONTRIBUTIONS

---

- Pieper, S., 81  
Polak, M., 60  
Ponsoda, V., 2  
Powell, J.C., 61  
Price, L., 61  
Priebe, S., 63  
Rabe-Hesketh, S., 10, 61  
Raîche, G., 62, 73  
Rausch, J.R., 62  
Raykov, T., 62  
Reckase, M.D., 39, 59, 63  
Reininghaus, U., 63  
Rich, C., 93  
Ricks, R.A., 64  
Rijmen, F., 64  
Ripley, B.D., 30  
Rizopoulos, D., 54  
Ro, S., 77  
Roberts, J.S., 64, 69  
Romeijn, J.-W., 3  
Rosopa, P.J., 65  
Roussos, L., 18, 65, 67, 84  
Roxbury, T., 4  
Ruggeri, M., 63  
Rupp, A.A., 41, 96  
Rust, J., 65  
Ryoo, J., 66  
Samejima, F., 66  
Sano, M., 66  
Satorra, A., 54, 67  
Schepers, J., 26  
Scholl, L.H., 67  
Seo, D.G., 66  
Seo, M., 67  
Seong, J., 86  
Serrano, D., 68  
Setzer, J.C., 68  
Sheng, Y., 64, 69  
Shigemasu, K., 53, 58, 69  
Shim, H.S., 69  
Shimizu, S., 70  
Shimizu, Y., 25, 40  
Shojima, K., 70  
Shrout, P.E., 42, 70  
Shu, L., 71  
Shyu, C.-Y., 71  
Sijtsma, K., 3, 12, 46, 97  
Sinharay, S., 71, 85  
Skrondal, A., 61, 72  
Slade, M., 63  
Smith, J., 72  
Smithson, M., 72  
Song, J., 73  
Sotaridona, L.S., 73  
Steinley, D., 73  
Straatemeier, M., 82  
Strawderman, W., 27  
Stucky, B.D., 68, 74  
Suk, H.W., 74  
Sun, R., 74  
Suzukawa, Y., 29  
Swartz, R.J., 75  
Swoboda, C.M., 38  
Takai, K., 75  
Takane, Y., 31, 32, 75, 76  
Takeshita, M., 29  
Tala, A., 76  
Tatsuoka, K., 15  
Telesca, D., 17  
ten Holt, J.C., 76  
Thayer, J.F., 81  
Thissen, D., 74, 76, 77  
Thompson, N.A., 77  
Thompson, V.M., 64  
Timmerman, M.E., 77  
Tofighi, D., 78  
Toyoda, H., 20, 29, 35  
Trendafilov, T., 80  
Trierweiler, T.J., 78  
Tsai, R.C., 78  
Tuerlinckx, F., 4, 49, 79  
Uher, R., 79  
Umehara, T., 79  
Unkel, S., 80  
Usami, S., 80  
van Assen, M.A.L.M., 12  
van der Ark, L.A., 46, 80, 97  
van der Leeden, R., 81  
van der Linden, W.J., 20, 39, 81  
van der Maas, H.L.J., 33, 50, 82  
van der Palm, D.W., 80  
van Duijn, M.A.J., 76  
Van Ginkel, J., 82  
Van Mechelen, I., 26, 82, 83, 88

## ABSTRACTS OF THE CONTRIBUTIONS

---

van Rijen, S., 3  
van Rosmalen, J., 83  
van Schuur, W.H., 83  
Verkuilen, J., 6, 72, 83  
Vermunt, J.K., 49, 84, 94  
Verschoor, A., 84  
Voge, N., 17  
von Davier, A.A., 44  
von Davier, M., 85, 92  
Walker, L., 89  
Walls, T.A., 73, 86  
Wang, A., 86  
Wang, Chun, 86  
Wang, Chunxin, 87  
Wang, L., 87  
Wang, T., 87  
Wang, W., 88  
Wang, Y., 88  
Wells, C.S., 24  
Wicherts, J., 3  
Wilderjans, T., 88  
Wilhelm, O., 41  
Willson, V.L., 74, 89  
Wilson, M., 6, 89  
Wise, S.L., 39, 89  
Wolf, A.N., 65  
Wood, P.K., 30  
Woods, C., 47, 90  
Wu, C., 4  
Wu, H., 90  
Wu, J.-Y., 90  
Wu, Y.-Y., 28  
Wyse, A.E., 91  
Xu, S., 91  
Xu, X., 91, 92  
Xu, Z., 8, 14  
Yamamoto, K., 92  
Yang, M., 92  
Yang, N., 93  
Yang, X., 16  
Yang-Wallentin, F., 93  
Yao, L., 27, 93  
Yates, P.D., 15  
Ye, F., 67, 94  
Yoon, M., 78  
You, W., 94  
You, X., 14  
Yu, H.-T., 94  
Yuan, K.-H., 49  
Yung, Y.-F., 94  
Zervoulis, K., 65  
Zhang, G., 95  
Zhang, J., 95  
Zhang, S., 87, 95  
Zhang, T., 96  
Zhang, Y.-P., 96  
Zhao, T., 8, 14  
Zheng, X., 89  
Zhu, Y.-F., 50  
Zijlstra, W.P., 97  
Zou, Y., 28  
Zu, J., 97