IMPS 2014 Madison Wisconsin

ABSTRACTS

79th Annual Meeting of the Psychometric Society
Madison, Wisconsin, USA

ABSTRACTS

Pre-Conference Workshops
July 21, 2014

Annual Meeting
July 22-25, 2014

Abstract Presentations:
All abstracts are listed in alphabetical order by first author.

Paper Presentations:
Authors have been invited to present between Tuesday, July 22 - Friday, July 25, 2014. Refer to the Schedule-At-A-Glance, in the 2014 IMPS Conference Program book for specific times and locations.

Poster Presentations:
Authors have been invited to be present during the Poster Session and Reception on Thursday, July 24, 2014. The session will take place between 4:30 p.m. - 6:30 p.m. in the Pyle Center Lee Lounge and AT&T Lounge.

A Method to Combine Trace Lines of Distractors in an Item Characteristic Chart

Poster Presentation

Takashi Akiyama, *Waseda University*
Hideki Toyoda, *Waseda University*
Norikazu Iwama, *Center for Japanese-Language Testing, The Japan Foundation*

An item characteristic chart is a tool to investigate the functioning of multiple-choice items. Drawing trace lines for distractors, in addition to a correct option, visualizes the propensity of incorrect responses of examinees and facilitates error analysis. However, it is not recognized widely as a criterion for verifying whether or not the propensity is sampling fluctuation. Furthermore, trace lines for incorrect options tend to have similar shapes and overlap with each other. This tendency often reduces the visibility of uniquely-shaped, thus informative, trace lines. One approach to this problem is to combine similar trace lines on the basis of hypothesis constructed from the error analysis to reduce the number of lines in the chart. Although this approach seems straightforward, no widely accepted criterion to evaluate the similarity of trace lines of distractors is available in the literature. In this poster presentation, we proposed a method to calculate AIC and BIC for item characteristic charts, and investigated its performance using real and simulated data. The results of this study indicate that the proposed method can verify the hypothesis derived from error analysis, and improve the visibility of item characteristic charts.

Estimation of Speed-accuracy Response Models: Scoring Rules for Adaptive Tests

Paper Presentation

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

The availability of RT information brings the opportunity for studying various forms of using such information. We aim to explore a third approach initiated by Maris and van der Maas (2012) where speed and accuracy are inserted in straightforward scoring rules. We derived psychometric models and their corresponding item information functions suitable for two different scoring rules: one that considers only positive scores, and one that considers both positive and negative scores. We found that these scoring rules can improve measurement precision by applying them to real data. As including RT in scoring improves measurement precision, it might be useful in enhancing the item selection algorithm. We investigate the use of response time (RT) in adaptive testing. Combining response and RT in scoring might be useful in 1) enhancing the item selection algorithm in item-level adaptive tests, and 2) creating easy routing blocks in multistage tests without losing precision. Implications of results are discussed.

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

Paper Presentation

Abdullah AL-Sadaawi, *National Centre for Assessment in Higher Education*

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

Application of the Posterior Predictive Check Method for Two-dimensional Data

Poster Presentation

Allison Ames, *University of North Carolina at Greensboro*
Terry Ackerman, *University of North Carolina at Greensboro*

This project investigates how prior specification influences the Posterior Predictive Check method for the 2-dimensional compensatory multidimensional IRT model. Model checking in the Bayesian paradigm encompasses the 1) sampling model and 2) quantification of the investigator's prior beliefs. Both of these model checking areas represents a way in which the Bayesian model can fail, requiring the Bayesian model to be evaluated individually for these separate sources of model failure. The Posterior Predictive Check (PPC) model check method addresses only the sampling model appropriateness, often studied using large sample sizes and noninformative priors so that the data drives Markov Chain Monte Carlo estimation rather than the priors themselves. However, the use of noninformative priors can often lead to improper posteriors and introduce bias into parameter estimates. To address the issue of prior specification under the PPC method, a simulation study will be conducted. Data will be generated according to the 2-d compensatory model, but fitted to a unidimensional 2PL model. Two test lengths (I= 10, 30 items) and three sample sizes (N= 500, 1000, 2000) will be investigated. Three prior specifications will be used: noninformative, accurate-informative, and inaccurate-informative. A variety of discrepancy statistics will be used to detect unaccounted for dimensionality.

Conditional Multidimensional Item Response Models for Polytomous Items

Paper Presentation

Carolyn Anderson, *University of Illinois, Urbana-Champaign*
Hsiu-Ting Yu, *McGill University*

Log-multiplicative association models (LMA), which are special cases of log-linear models, have been derived as models of manifest probabilities based on multidimensional latent variables models for multi-category response data. Since discussion of between multiple derivations is lacking, we draw connections between multiple derivations with an emphasis on those pertaining to item response modeling. In particular, we extend Anderson and Yu (2007) who studied LMA models as uni-dimensional IRT models for dichotomous items to the case of multi-dimensional latent variables for polytomous items. This research also extends Hessen's (2012) conditional multinomial partial credit models to more general IRT models. The empirical studies presented here demonstrate that LMA models perform extremely well compared to standard MIRT models in terms of model fit to data, parameter recovery, and estimation of values on the latent variables. Some example applications using data from a study on bullying (Espelage & Holt 2010).

Hypothesis Testing of Equating Functions in IRT Observed-Score Kernel Equating

Paper Presentation

Bjorn Andersson, *Uppsala University*

Item response theory (IRT) can be utilized in observed-score equating by generating the score distributions implied by the IRT models and calculating the approximated continuous distribution functions of the tests. Then an equipercentile equating is conducted. The IRT observed-score equating procedure has the advantage over observed-score equating with log-linear models in providing covariance matrices for the equating function which are of full rank. A Wald test for hypothesis testing of linear hypotheses of equating functions in IRT observed-score kernel equating is proposed which allows for testing hypotheses regarding all score points simultaneously. The hypothesis test is applied to the non-equivalent groups design using post-stratification equating and chain equating and also to the method of local equating.

Multiple Correspondence Analysis for Test Items

Poster Presentation

Sayaka Arai, *The National Center for University Entrance Examinations*

Multiple correspondence analysis (MCA) is a method for quantifying categorical multivariate data. It represents data as clouds of points in a low-dimensional space. In this study, application of MCA to academic examination data is considered. Academic examination data is usually analyzed using classical test theory (CTT) or item response theory (IRT). These test theories characterize each test item by its item statistics such as item difficulty and item discrimination. On the other hand, MCA represents the

relationships of alternatives within each test item. Therefore, MCA may provide useful information for item analysis. An empirical study using practical reading test data is done. Items that have alterations, which differ only in the alternatives, are focused on and the effects of the changes in alternatives are analyzed. The analysis using MCA is compared to that of other methods.

A Dynamic Item Response Model with Time-Dependent Autoregressive Item Difficulties

Paper Presentation

Ethan Arenson, *National Board of Osteopathic Medical Examiners*
George Karabatsos, *University of Illinois at Chicago*
Stephen G. Walker, *University of Texas at Austin*

Many IRT researchers have explored different ways of modeling parameters as dynamically changing over time. Virtually all of the research on dynamic IRT models is based on either longitudinal or dynamic GLM modeling strategies. However, it is known that longitudinal and dynamic linear modeling strategies are not fully satisfactory, because these models assume that: (i) observed values occur at discrete and equidistant time intervals; and (ii) changes over time with respect to the latent trait being measured are linear. Therefore, we introduce a dynamic change-point IRT model, which has the advantages of added modeling flexibility, particularly with respect to the aforementioned points. The new model, which we characterize as an infinite mixture model with a time-dependent mixture distribution for the item parameters can handle the analysis of either dichotomous or polytomous item responses. One can easily extend this model in the following ways: (i) to conduct multilevel analyses, for example, in order to handle the analysis of data where examinees are nested within schools; (ii) to analyze dynamically changing examinee abilities; and (iii) to analyze multidimensional abilities. We illustrate this model through the analysis of data from the Trends in Mathematics and Science Survey (TIMSS).

Multidimensional Item Response Theory Models with Multivariate Skew Normal Latent Trait Distributions Under the Centered Parameterization: Bayesian Parameter Estimation, Structural Selection and Model Fit Assessment

Poster Presentation

Caio Lucidius Naberezny Azevedo, *University of Campinas*
Juan L. G. Padilla, *University of Campinas*
Victor H. Lachos, *University of Campinas*

Multidimensional IRT (MIRT) models are quite useful to analyze data sets involving multiple skills or latent traits, which is the case in many applications. However, most of the work available in the literature consider the usual assumption of a multivariate (symmetric) normal distribution to model the latent traits, do not handle a multiple groups framework (few groups with a lot of subjects in each one), consider only a limited number of model fit assessment tools, do not investigate measurement instrument dimensionality in a detailed way, and handle the model nonidentifiability in a non-trivial way. In this work we propose a MIRT multiple group model with multivariate skew normal distributions for modeling the latent trait of each group under the centered parameterization presenting simple conditions for the model identification. A full Bayesian approach for parameter estimation, structural selection (model comparison and determination of the measurement instrument dimensionality) and model fit assessment is developed through MCMC algorithms. The developed tools are illustrated through the analysis of a real data set related to the 2013 first stage of the University of Campinas admission exam.

IRT Parameter Linking: Standard Errors of Linking and Optimal Linking Design

Paper Presentation

Michelle D. Barrett, *CTB/McGraw-Hill*
Wim J. van der Linden, *CTB/McGraw-Hill*

As linking functions for item response models are derived from estimated model parameters, their error automatically propagates into linking error. Asymptotic standard errors (ASEs) of linking estimators may be approximated by applying a first-order multivariate delta method to the linking function. We present expressions for the ASEs for a new type of estimator of linking function for various common-item and common-person designs that proportionately reduces the influence of common items or persons with larger parameter estimation error. The ASE of this estimator will be compared to those of current mean/mean and mean/sigma methods for multiple empirical examples from recent linking studies. We will then demonstrate how using this approach facilitates optimal linking design, in which the set of common items between two independently calibrated test forms may be optimally selected to minimize linking error.

Individual Differences in Cognitive Modelling: A Bayesian Hierarchical Mixture Approach

Paper Presentation

Annelies Bartlema, *KU Leuven*
Michael Lee, *University of California, Irvine*
Ruud Wetzels, *University of Amsterdam*
Wolf Vanpaemel, *KU Leuven*

Unlike psychometric modeling, cognitive modeling has aimed at extracting commonalities between people rather than their differences. However, ignoring individual differences can lead to distorted conclusions. In this talk, I will discuss different types of individual differences: continuous differences, where individuals smoothly vary around some central tendency; and discrete differences, where there are more fundamental differences between people. I will demonstrate the potential of using a Bayesian hierarchical mixture approach to model both types of individual differences simultaneously. Mixture components can be used to identify latent groups of subjects, who use qualitatively different cognitive processes, or very different parameterizations of the same cognitive process, while hierarchical distributions can be used to capture more minor variations within each group. The Bayesian hierarchical mixture approach can be applied to problems that are typically conceived of as a problem of parameter estimation, and to problems that are traditionally tackled from a model selection perspective. I will demonstrate the flexibility and wide applicability of the hierarchical mixture approach to modeling individual differences in two illustrative applications, involving how people learn and use categories.

Covariate-free and Covariate-based Reliability

2013 Career Award Keynote Address

Peter Bentler, *UCLA*

Classical test theory contains parameters such as reliability coefficients that can be population specific, but methods for judging the degree of invariance of such quantities across contexts are not well developed. When a given scale has scores that are described as having a given reliability, it is hard to know how generalizable or subject to conditions such a result may be.

Reliability generalization meta-analysis has provided the main methodology for addressing this issue (e.g., Vacha-Haase & Thompson, 2011). A single-sample methodology is proposed here for evaluating the effects of covariates on reliability. Coefficient alpha, 1-factor reliability, model-based reliability, the greatest lower bound, and quantile lambda-4 reliability are partitioned into additive covariate-free and covariate-dependent reliability coefficients whose sum is the defined reliability measure. In this setting, high reliability generalization occurs when, for a meaningful selection of possibly confounding covariates, covariate-based reliability is trivially small.

Visualizing and Clustering Process Data Sequences from a Simulation-Based Task

Paper Presentation

Yoav Bergner, *Educational Testing Service*
Zhan Shu, *ETS*
Jiangang Hao, *ETS*
Mengxiao Zhu, *ETS*
Alina von Davier, *ETS*

In the third analysis in this series, challenges of visualization and clustering are explored with respect to sequence data from the Wells scenario-based task. Visualization issues include representing progress towards a goal and accounting for variable-length sequences. Clustering issues focus on external criteria with respect to official scoring rubrics of the same sequence data. The analysis has a confirmatory flavor; the goal is to understand to what extent clustering solutions align with score categories. It is found that choices related to data preprocessing, distance metric and external cluster validity measures all impact agreement between cluster assignments and scores. This work raises key issues about clustering of educational data, especially in the presence of multidimensionality. Different clustering protocols may lead to different solutions, no one of which is uniquely best.

Improved Measurement of Noncognitive Constructs with Anchoring Vignettes

Paper Presentation

Jonas P. Bertling, *Educational Testing Service*
Patrick C. Kyllonen, *Educational Testing Service*

In PISA and other large-scale assessments and in psychological surveys, indices capturing noncognitive factors (e.g., motivation, personality) are often based on self-report items with the vaguely defined response categories "strongly agree" to "strongly disagree". Multiple problems have been reported for indices based on this response format, such as lack of cross-cultural comparability, response category DIF, and proneness to response styles. In this paper, we propose an alternative scoring approach for Likert-type questionnaire items that is based on anchoring vignettes. We present validity data based on PISA 2012 that indicates that anchoring vignettes can substantially improve international comparisons. Anchoring provides an alternative metric for comparisons that reduces response category DIF. Effects on three different level (student, school, country) are distinguished. Best practices for using anchoring vignettes with nonparametric scaling models and frequency diagnosis including analysis of ties and order violations will be described. The paper discusses implications of these findings for the assessment of noncognitive constructs in K-12, higher education and workforce populations in general and provides a rationale why these findings should change how we think about measuring noncognitive constructs, especially in cross-national and multi-cultural populations.

Measuring Change for Individuals and Groups: Evaluating Standardized Growth Norms

Paper Presentation

Joseph Betts, *Pearson*

The evaluation of repeated measures data using a longitudinal design has become common in education and the health sciences. In education continuous progress-monitoring has emerged as a major methodology for tracking student progress across interventions and end-of-year summative testing to monitor progress over an entire academic career. In the area of health sciences, continued follow-up measurements have been greeted as an important method for evaluating the course of treatment. These types of outcome monitoring have been utilized not just to make decisions about individuals, but aggregate individual information informs judgments about larger institutions. This research evaluates the issues related to the reliability of changes scores both at the individual and aggregate level. The method of Standardized Growth Norms (SGN) will be presented as a potentially useful method of making inferences about growth for individuals by using normative estimates of growth. Evidence will be provided to show that aggregating SGN provides an unbiased method for making inferences about organizational change that relies on the aggregated data. Additionally, this methodology will be discussed in the context of an alternative model for studying the effect sizes of treatment interventions rather than relying on a single effect size estimate normally reported.

An Extension of the CORANOVA Method for Correlated Correlations

Paper Presentation

Warren B. Bilker, *University of Pennsylvania*
Colleen M. Brensinger, *University of Pennsylvania*
Ruben C. Gur, *University of Pennsylvania*

Standard procedures for testing the homogeneity of independent Pearson correlations, such as the Fisher's Z-test, are not applicable when the correlations are themselves correlated, as when there is a common variable being correlated. Suppose measurements of brain activation are taken in three regions, in left and right hemispheres, before and during a verbal memory task, measured on a scale, V. The relative pre-post percent changes in brain activity in each region and hemisphere, C_RH, are estimated. Of interest are the correlations rho(V,C_RH) for each region and hemisphere. V is common to all correlations, which are thus called "correlated correlations". A testing procedure for the homogeneity of these correlations between hemispheres, regions, and two or more independent groups (e.g. diagnostic group) and all two and three way interactions of these between and within factors is presented. In general, two within and one between factor are considered. The method is analogous to a two-way ANOVA with interaction, except for homogeneity of correlations, rather than means. The null distributions for testing each effect is estimated via permutation based methods, avoiding asymptotic distributional assumptions and enhancing applicability to sample sizes and non-normal data. The power of main and interaction effects is assessed via simulations.

A Joint Model Approach for Longitudinal Data with a Binary Outcome

Paper Presentation

Brenden Bishop, *The Ohio State University*

Joint Models are two stage statistical models which specialize in the analysis of predictors and responses that are of disparate forms. There is an extensive literature on Joint Models within biostatistics, but the models are underutilized in psychology. The present research illustrates the benefit of using a Joint Model for longitudinal data in which the ultimate research question is binary. The data are from a classic dataset by Williamson, Appelbaum, and Epanchin (1991) in which scholastic achievement tests were given across eight years to schoolchildren in North Carolina. The Joint Model is applied by taking characteristics from subject specific response curves and utilizing them as latent predictors for a logistic link function which joins the longitudinal measures to the binary covariate. Advantages of the Joint Model are discussed.

Multi-trait Multi-method Models for Interchangeable vs. Structurally Different Methods: Consequences of Specification Error

Paper Presentation

Jacob Bishop, *Utah State University*
Christian Geiser, *Utah State University*
Ginger Lockhart, *Utah State University*

Confirmatory factor analysis (CFA) models are frequently applied to examine the convergent validity of scores obtained from multiple raters or methods in multi-trait multi-method (MTMM) investigations. Many applications of CFA-MTMM are plagued by improper or close-to-improper solutions, in which one or more method factors show zero or non-significant variance estimates. Eid et al. (2008) distinguished between MTMM measurement designs with interchangeable (randomly selected) versus structurally different (fixed) methods and showed that the type of design logically determines which measurement model should be used to analyze the data. We hypothesized that some problems commonly seen in applications of CFA-MTMM would arise if researchers incorrectly applied models for interchangeable methods to data generated by structurally different methods and vice versa (specification error). Using simulations, we tested the hypothesis that the common problem of one method factor having a negative or non-significant variance may be related to specification error. We found that applying models designed for interchangeable methods to data generated from structurally different methods indeed led to a higher proportion of solutions in which at least one method factor was unstable. These findings have implications for other SEM models, including mediation models, nested or bifactor models, and latent state-trait models.

When Do Response Styles Distort Attitude Measurements? A Comparison of Different Response Formats

Paper Presentation

Ulf Bockenholt, *Northwestern University*

Focusing on response-style (RS) effects in Likert ratings, this presentation makes two contributions: First, I review two positions in the literature ("applied camp" and "method camp") with opposing views about the importance of RS effects in ratings. The "applied camp" argues that RS effects can be ignored, whereas the "method camp" stresses that RS have detrimental effects. To investigate the merits of these positions, I present the results of a simulation study of a response-tree model that can account for RS effects. This study shows that the Pearson correlation coefficient is not sensitive to RS effects in ratings which explains the position of the "applied camp" that RS effects are not important. Second, I show that response-tree models are useful in analyzing the effects of different rating-scale layouts. An empirical study comparing three different response formats demonstrates that these response formats yield similar trait information but elicit different RS types and levels.

Probabilistic Reasoning: How Should a Diagnostic Test be Evaluated?

Paper Presentation

Ehsan Bokhari, *University of Illinois at Urbana-Champaign*

The area under the receiver operating characteristic (ROC) curve (AUC) is a widely-used measure for assessing diagnostic test performance. Despite this popularity, the AUC is not a good measure of a test's overall performance when used with populations having differing base rates for the characteristic being assessed. We present several reasons and accompanying illustrative examples as to why the AUC measure is misleading and subject to bias. We suggest the use of the positive and negative predictive values to supplement or replace the AUC; these incorporate base rate information and indicate whether a test will actually outperform prediction that simply uses base rate information.

Evaluation of a New Cluster Analytic Approach in Exploratory Structure Detection

Paper Presentation

Stella Bollmann, *Ludwig Maximilians University*

Within the framework of evaluating test items, exploratory factor analysis (EFA) is a widely accepted standard for exploration of test structures and detection of items that have a similar content. Some researchers (Bacon, 2001; Hunter, 1973; W. Revelle, 1979; Schweizer, 1991) suggested a less popular method for the initial exploratory assessment of psychological tests, which does not follow a latent variable approach - Cluster Analysis (CA). Nevertheless, none of the hierarchical CA methods that have been tested so far can compete with the performance of EFA (Bollmann et al., under review). Furthermore, for item clustering there is no standardised method to determine the number of clusters that is known to perform well. In this paper, a new k-means approach for clustering of items and a new approach for assessment of dimensionality is tested. This is done in a traditional simulation study in which data is generated according to a factor model and different parameters are manipulated. Additionally, the methods are tested in a real world simulation in which subsamples of a real and large data set are analysed. Results suggest that k-means clustering is a useful alternative to EFA and that it performs better than hierarchical clustering methods.

Time-varying Networks: Applying a Novel Method to Interpersonal Relations

Paper Presentation

Laura Bringmann, *KU Leuven*

Interpersonal relations are often unstable over time. A common way to deal with non-stationary time series involving relations that change over time is to make the data stationary by transforming them (e.g., differencing). However, transforming to stationarity often leads to an information loss, including loss of potentially interesting information such as the fact that the interpersonal relations are changing, and the shape of this change. In this paper, we examine varying coefficient vector autoregressive (VCVAR)

models, where both the means and transition coefficients are allowed to change smoothly over time. This smooth change is modeled using penalized regression splines. We will illustrate this approach by applying it to a network of the variables measured in six astronauts who were in isolation for 500 days as part of the Mars-500 project. This results not in a single static network, representing time constant dynamics, but in a network with time varying dynamics.

Sensitivity of Fit Indices to Test for Cluster Bias in Multilevel Data

Poster Presentation

Qian Cao, *Texas A&M University*

Mark H. C. Lai, *Texas A&M University*

Multilevel structural equation modeling (SEM) is appropriate for testing measurement invariance across clusters (Jak, Oort, & Dolan, 2013; Muthen, 1990; Rabe-Hesketh, Skrondal and Pickles, 2004), as it overcomes the inefficiency of multi-group factor analysis (MGFA) when the number of groups is large (e.g., larger than 30). Under the multilevel SEM framework, the purpose of this Monte Carlo simulation study is to investigate the performance of model fit indices (with conventional cutoff values) for detecting cluster bias based on the methods illustrated by Jak and colleagues in 2013. Five design factors (i.e., cluster size, number of clusters, number of non-invariant intercepts, percentage of clusters with non-invariant intercepts, and ICC) are manipulated. From preliminary results we found that RMSEA, CFI, and ΔCFI are virtually insensitive to measurement invariance in the between-level (i.e., cluster bias), which is consistent with previous studies (i.e., Hox, 2010; Ryu & West, 2009). The $\chi 2$ test for exact fit is the most powerful (although with inflated Type I error rate), and the $\chi 2$ difference test is a more conservative test but with better Type I error control. The SRMR-Between works poorly when cluster size is small (i.e., 5). We recommend researchers stay with the $\chi 2$ test when testing cluster bias and exert cautions when they do not have a large sample size.

A New Constrained Approach for Latent Class Modeling of Response Styles in Item Responses

Paper Presentation

Claus H. Carstensen, *University of Bamberg*

Eunike Wetzel, *University of Tuebingen, Germany*

Responses to personality questionnaires do not inform about a latent trait intended to measure only, but are subject to depend on additional person characteristics as well, i.e. response styles. Mixture distribution item response models have been used to model response styles, e.g. the tendency for extreme responses (Rost, Carstensen & von Davier, 1997). The interpretation of latent classes as different response styles was quite obvious, however not formally defined. For the model presented in this paper, a constraint on the location parameters between latent classes of a mixed partial credit model is introduced to fix the latent trait being measured to be identical between classes. If this model holds for more than one class, a response style can be assumed. Furthermore the latent trait estimates between classes may be compared. We illustrate the model with example data from the NEO-PI-R.

The Longitudinal Hierarchical Rater Model for Evaluating Changes in Traits

Paper Presentation

Jodi Casabianca, *The University of Texas at Austin*
Brian Junker, *Carnegie Mellon University*

The hierarchical rater model (HRM) is a multilevel item response theory model for multiple ratings of responses, behavior and performance. This research presents an extension of the HRM made to analyze ratings assigned within longitudinal designs. In the extension, an additional level is added to the hierarchy for time, so that latent traits can be estimated at more than one time point. The presentation will discuss the statistical framework and extension, as well as simulation study results that evaluate the extension for feasibility under various conditions.   A Monte Carlo simulation study was used to determine how the longitudinal HRM performs under various conditions for a 5-item test and 500 simulees. We varied the number of time points, (3, 6, 9), the number of raters, (2,6), and combinations of individual rater bias and variability. The design is fully crossed so that each combination of factors is evaluated. The model was fitted using MCMC estimation in Jags. Preliminary results show estimated traits increasing with time and well-recovered rater parameter estimates of bias and variability. In our presentation we will report on the full study results and discuss situations for which the longitudinal HRM is a feasible model for rated data in psychology and education.

Graphical Utilities for Semiparametric Structural Equation Models in R

Poster Presentation

R. Philip Chalmers, *York University*
Jolynn Pek, *York University*

Linear structural equation models (SEMs) are commonly used to estimate relationships between latent variables. Recent developments in non-linear SEMs involving a semi-parametric model (SPM) allow for the recovery and detection of potential non-linearity (Bauer, 2005; Bauer, Baldasaro, Gottfredson, 2012; Pek, Sterba, Kok & Bauer, 2009; Pek, Losardo, & Bauer, 2011). Such SPMs are difficult to implement, and require considerable post-processing of results. Our R package 'plotSEMM' processes results directly from Mplus and generates smoothed latent regression functions recovered by SPM, including their non-simultaneous confidence intervals and simultaneous confidence bands. The package also includes a line finding algorithm that implements an informal test to diagnose latent non-linearity. For users less familiar with the R programming environment, we also provide an interactive web-based graphical user interface implemented with the R package 'shiny' (RStudio & Inc., 2014).

Gender-Based Differential Item Functioning in PISA 2012 Science Assessment: Evidence from 68 Economies

Poster Presentation

Jehanzeb R. Cheema, *University of Illinois at Urbana-Champaign*

This study investigated gender-based differential item functioning (DIF) in science literacy items included in the Program for International Student Assessment (PISA) 2012. Prior research has suggested presence of such DIF in large scale surveys. Our study extends the empirical literature in two ways. First, unlike many previous studies we test for both uniform and non uniform DIF in science assessment items. Second, we examine gender-based DIF differences at the country level in order to gain a better overall picture of how cultural and national differences affect occurrence of DIF. Our statistical results indicate existence of widespread gender-based DIF in PISA with estimates of percentage of potentially biased items ranging between 2% and 44% (M = 16, SD = 9.9). Our reliance on nationally representative country samples allows these findings to have wide applicability.

A Two-step Bayesian Propensity Score Approach for Multilevel Observational Studies

Paper Presentation

Jianshen Chen, *University of Wisconsin-Madison*

Despite the fact that observational studies with nested data structure are commonly seen, only a limited amount of studies explore the use of multilevel modeling in nonrandomized designs for causal inference. To the best of our knowledge, Bayesian propensity score approaches for multilevel observational studies have not yet been studied in the literature, but have the unique benefit of naturally accounting for parameter uncertainty. This paper extends the two-step Bayesian propensity score approach into multilevel settings and provides a practical Bayesian propensity score approach for making causal inference in multilevel observational studies while accounting for uncertainty in both propensity score and outcome. The performance of the proposed Bayesian approach is examined for different matching strategies, model specifications, prior information, level-one and level-two sample sizes, intra-class correlations and propensity score methods via two comprehensive simulation studies. Results showed that the proposed approach offers less biased causal effect estimates and more accurate uncertainty estimates compared to models that ignore the multilevel structure. Overall, a Bayesian random intercept and slopes propensity score model for within-cluster matching strategy is recommended. When there is little evidence of omitted cluster-level covariates, a Bayesian multilevel model with across-cluster matching can provide as good treatment effect and variation estimates as within-cluster matching.

Calibrating Automated Scoring Models Using Machine Learning Methods

Paper Presentation

Jing Chen, *Educational Testing Service*

E-rater® is an automated essay scoring system used in Educational Testing Service. In the current practice, the score that e-rater gives to an essay is generated by a linear combination of a set of feature scores of the essay developed using natural language processing technology. Limitations are inherent in the over simplified assumptions of linear models. Algorithms have been developed in the machine learning community to address the multidimensional classification and regression problem. This study aims at building e-rater models using several machine learning algorithms, such as Support Vector Machine (SVM), Random Forest, and k-Nearest Neighbor Classification. The results from this study suggest that models based on machine learning algorithms outperform linear regression models in predicting human raters' scores. Among the three machine learning algorithms that we tried in this study, SVM based models have the highest agreement between human and e-rater scores. Comparing to linear regression based models, SVM based models significantly improved the agreement between human and e-rater scores at the ends of the score scale. In addition, the high correlation between SVM based e-rater scores and external measures such as examinees' scores on the other parts of the test provide validity evidence for SVM based e-rater scores.

Evaluation of the Psychometric Properties of the Basic Literacy Test for University Students

Paper Presentation

Po-Hsi Chen, *National Taiwan Normal University*
Hsin-Ying Huang, *National Chengchi University*
Po-Wei Li, *National Taiwan Normal University*
Yu-Shin Chen, *National Taiwan Normal University*
Tai-Ting Yeh, *National Taiwan Normal University*
Chun-Yu Hsu, *National Taiwan Normal University*
Su-Shao Zu, *National Taiwan Normal University*

Investigation of literacy of the students was an important issue in recent years. This three-year research project has been initiating research to develop valid, reliable, and practical assessments of basic literacy for university students in Taiwan. The basic literacy assessment is comprised of nine literacy domains, including communication and collaboration, aesthetics, information literacy, lifelong learning, career, leadership, problem solving, social concern and citizenship, and scientific thinking. Besides, 2~3 literacy testlets are organized by each domain. The purpose of this study is to evaluate the psychometric properties of this assessment. A total of 10958 students from 20 universities participated in this study. Partial credit model was applied to analyze the standard errors of the estimated person parameters for reliability of the basic literacy assessment. Results demonstrated that reliabilities ranged from .61 to .71 in nine literacy domains. Furthermore, validity evidence was obtained by grouping students based on Holland's hexagon model. The performance of different types of students in basic literacy assessment was consistent with Holland interest patterns. All the evidence collected supports the high degree of reliability and validity of the basic literacy assessment. The basic literacy assessment was recommended to apply extensively.

A General SEM Framework for Integrating Moderation and Mediation

Poster Presentation

Shu-Ping Chen, *National Chengchi University*

Modeling the combination of moderating and mediating effects is a significant issue in the social and behavioral sciences. Edwards & Lambert (2007) presented an analytical framework for integrating the two effects through moderated path analysis. However, their framework is not equipped to deal with multiple measures of constructs and measurement error, making it largely unsuitable for psychology research. In this article, we further generalize Chen & Cheng (2014) to develop a framework which supports latent variable versions of all the models from Edwards & Lambert (2007), including the three-way interaction models (i.e., First and Second Stage Moderation Model and Total Effect Moderation Model). Similar to Chen & Cheng (2014), the constraint specification procedure is matricized and partitioned to fit into the advanced framework. The usage and validity of the procedure is demonstrated with several simulated data set examples via OpenMx. The current study represents a further step forward in the development of the constrained approach on both theoretical and practical grounds.

Q-matrix Estimation for Diagnostic Classification Models via a Regularized Likelihood Approach

Paper Presentation

Yunxiao Chen, *Columbia University*

Diagnostic classification models have recently gained prominence in educational assessment, psychiatric evaluation, and many other disciplines. Central to the model specification is the so-called Q-matrix that provides a qualitative specification of the item-attribute relationship. In this paper, we propose a computationally affordable method for the estimation of a Q-matrix based on the regularized maximum likelihood. This regularized estimator can be computed by means of a combination of the expectation-maximization algorithm and the coordinate descent algorithm. The proposed method can be applied to a large class of diagnostic classification models.

Classification Estimation of the DINA model

Paper Presentation

Yuehmei Chien, *Pearson*

Ning Yan, *Independent Consultant*

Chingwei David Shin, *Pearson*

Recently, the diagnostic classification models (DCMs), which aim to determine mastery or non-mastery of a set of attributes, have drawn much attention from practitioners. Among many DCMs, the deterministic inputs, noisy-and-gate (DINA) model is popular due to its simplicity and intuitive interpretation, and has been operationally applied to educational achievement assessment. In practice, the classification of mastery or non-mastery of each attribute is determined on the probability that is estimated based on a set of observed responses on the test (For example, an estimated probability of .91 indicates there is 91% of the chance that the latent class of that specific attribute is mastered). To obtain the classification results, a threshold of a probability needs to be predefined and then used to judge mastery or non-mastery status. It is not uncommon that a .5 threshold is adopted in real assessment; however, using a .5 threshold actually is flawed. This study first describes the reason that using .5 as a threshold is flawed. Then, the impact of different threshold values on classification accuracy is discussed using real data through simulation. Last, a simulation-based approach is proposed to determine an optimal threshold.

Joint Maximum Likelihood Estimation for Cognitive Diagnosis Models

Paper Presentation

Chia-Yi Chiu, *Rutgers, The State University of New Jersey*
Yi Zheng, *University of Illinois at Urbana-Champaign*
Robert Henson, *University of North Carolina, Greensboro*

Current methods for fitting cognitive diagnosis models (CDM) to item responses include Expectation Maximization (EM) algorithms used to formulate and maximize the marginal likelihood and Markov chain Monte Carlo (MCMC) techniques used to find the mode of the posterior probability distribution of a parameter. Joint maximum likelihood estimation (JMLE) is another alternative to estimate item parameters and examinees' attribute patterns. Unfortunately, JMLE has not been successfully implemented due to the potential inconsistency of parameter estimators, despite its attractive advantage of having simple likelihood functions. In this study, we propose a JMLE method that overcomes the theoretical deficiency that traditional JMLE suffers by using classification results obtained from the nonparametric classification (NPC; Chiu & Douglas, 2013) method as initial input. Two asymptotic consistency theorems for item parameter estimators are presented and proved. The proposed algorithm is empirically evaluated with simulation studies and an application to real data. The results show that the proposed JMLE method can effectively remedy the issues of existing estimation methods and estimate the item parameters and examinees' attribute patterns with high efficiency and accuracy.

Simultaneous Clustering of Parallel Factor Analysis

Paper Presentation

Ji Yeh Choi, *McGill University*
Heungsun Hwang, *McGill University*
Hsiu-Ting Yu, *McGill University*

Parallel factor analysis (PARAFAC) (Carroll & Chang, 1970; Harshman, 1970) is a useful decomposition method for three-way data. It has been combined with cluster analysis in either a sequential or unified manner in order to investigate potential cluster-wise characteristics inherent to the data (e.g., Rocci & Vichi, 2005). These extant approaches are geared only for partitioning entities in a single mode at a time. Thus, they often hardly show how a subset of entities in one mode is jointly associated with a subset of entities in another mode. In this paper, we propose a general approach to combining PARAFAC and cluster analysis in a unified manner. In the proposed approach, k-means is applied to classify PARAFAC components of each mode in such a way that a subset of entities in each mode is grouped simultaneously into a distinctive joint cluster. The approach provides cluster memberships of each mode's entities that belong to a joint cluster. The memberships are of help in relating a subset of each mode's entities to one another more clearly. We conduct a simulation study to evaluate the performance of the proposed approach. We also apply the approach to real data to demonstrate its empirical usefulness.

Relationship between Error Variance and Scaling Property for Student Growth Percentiles

Poster Presentation

Jinah Choi, *University of Iowa*
Won-Chan Lee, *University of Iowa*
Catherine Welch, *University of Iowa*
Stephen Dunbar, *University of Iowa*

Student Growth Percentiles (SGPs) methodology is one metric for indicating students' academic growth and gives norm-referenced growth information with their academic peers (Betebenner, 2009). The metric uses quantile regression as the statistical foundation. Also, the Standards (AERA/APA/NCME, 1999) recommended that Conditional Standard Errors of Measurement (CSEM) should be included for the accuracy or consistency of the metric when reporting the scale score. However, there is little study on errors for SGPs despite the increase of usage of the metric. In previous research on estimating CSEM for SGPs under a binomial model, it is found that the number of possible score points on SGP is much more than the number of items on the test, so it may lead to large SEMs at some score ranges (Choi, Lee, Welch, & Dunbar, 2014). Thus, this study investigates the effect of the units of quantile regression on the CSEMs for SGPs by using real longitudinal data. Since the unit plays a role in determining the characteristics of the SGP, this research is informative about the relationship between the error variance and the scaling properties of SGP.

A Simulation Study for Metric Identification in Mixture IRT Models

Paper Presentation

Youn-Jeng Choi, *University of Georgia*
Allan S. Cohen, *University of Georgia*

Mixture IRT models are being used with increasing frequency to explore the latent structure in test data. These models provide a tool for potentially better understanding how, and why, examinees give particular responses to test items. The metrics of the different latent classes need to be the same in order to compare response propensities. Lack of a common metric can result in failure to accurately measure differential performance between latent classes. The purpose of this study is to explore the performance of three different constraints for establishing a common metric in mixture IRT models: (1) equating using anchor items, (2) person centering, and (3) item centering (the mean of item diffculty parameters is set to zero). A simulation study will be presented to examine the performance of these constraints for mixture IRT models. The design of the simulation study includes the three types of constraints, two test lengths (20- and 40-items), two sample sizes (600 and 2,400 examinees), and three different numbers of latent groups (1-, 2-, and 3-groups) for the three dichotomous mixture IRT models. Twenty replications will be simulated.

Dynamic Systems Modeling in the Social and Behavioral Sciences

Invited Speaker

Sy-Miin Chow, *Pennsylvania State University*

From difference scores to confirmatory dynamic/longitudinal models, the study of change remains a central question of interest to researchers in the social and behavioral sciences. In the realm of dynamical systems modeling, the last decades have evidenced a gradual shift from heavy reliance on exploratory techniques to confirmatory approaches of studying dynamic processes via model fitting. Differential equation models provide a direct representation of change processes while allowing the data to be irregularly spaced – a common feature of data collected from ecological momentary assessment studies. In this talk, several existing approaches for fitting ordinary linear and nonlinear differential equation models with random effects to irregularly spaced data are considered. The use of functional data analysis approaches for model explorations, derivative estimation and model-fitting in a multi-step approach will be illustrated and compared with other single-step approaches for fitting ordinary differential equations.

Adapting Latent Profile Analysis to Semiparametrically Capture Individual Differences in Change over Time

Poster Presentation

Veronica T. Cole, *University of North Carolina, Chapel Hill*
Daniel J. Bauer, *University of North Carolina, Chapel Hill*

Most longitudinal methods used in psychology tend to make two assumptions about the nature of change over time: (1) that the shape of change follows a pre-determined functional form, and (2) that this same functional form applies to all individuals in the sample. The current report focuses on longitudinal latent profile analysis (LLPA), an exploratory mixture-based method which allows for the relaxation of these assumptions. In particular, the utility of this framework is explored in the context of individual-level inference. LLPA, as well as an extension thereof with a random effect, can be used to predict individual data points, and to assess individual differences at the person level. LLPA is compared to parametric mixtures (such as growth mixture models and semiparametric growth models) as well as latent curve approaches in its ability to recover individual trajectories in an empirical dataset, the NLSY97. Model-implied trajectories generated using LLPA are shown to provide a closer fit to individual data points than those generated from traditional methods. Bootstrapping-based and graphical methods for assessing uncertainty around individual predictions are also further investigated. By combining these techniques with LLPA, researchers will have a novel set of tools for making inferences at the individual level.

Speed-accuracy Response Models

Paper Presentation

Frederik Coomans, *University of Amsterdam*
Gunter Maris, *CITO - University of Amsterdam*
Han van der Maas, *University of Amsterdam*

The advent of large-scale computerized testing unlocked a new source of response data: response times. In a recent paper (Maris & van der Maas, 2012) a scoring rule for time limit tasks, based on both response accuracy and response time is proposed and the corresponding class of speed-accuracy response models is developed. In these models the score is a sufficient statistic for the person ability as well as for the item difficulty. We extend this class of models in two ways: by also considering a scoring rule that applies to open questions and by splitting the item parameter in a piece for which only the accuracy is sufficient (the difficulty) and a piece for which only the response time is sufficient (the time intensity). Our models differ from the ones developed in (van der Linden, 2007) in that they have just one person parameter for which the score is a sufficient statistic. In this way we explicitly keep a speed-accuracy trade-off into the models. We find that the time intensity parameter can be interpreted as a discrimination parameter and test our models using data from the computerized adaptive testing environment `Maths Garden' (Klinkenberg, Straatemeier & van der Maas, 2011).

New Robust Scale Transformation Methods in the Presence of Outlying Common Items

Paper Presentation

Zhongmin Cui, *ACT, Inc.*
Yong He, *ACT, Inc.*

Common items play an important role in IRT true score equating under the common-item nonequivalent groups design. Inconsistent item parameter estimates among common items can lead to large bias in equated scores. Current methods extensively focus on detection and elimination of outlying common items, which may lead to inadequate content representation of common items. To remove the impact of inconsistency in item parameter estimates while maintaining content representativeness, we proposed two robust scale transformation methods based on two weighting methods, the Area-Weighted method and the Least Absolute Values method. Results from two simulation studies indicate that the robust methods generally performed as well as the Stocking-Lord method in the absence of outlying common items and, more importantly, outperformed the Stocking-Lord method when a single outlying common item was simulated.

Applications of Sequential IRT Models to Cognitive Assessments

Paper Presentation

Steven A. Culpepper, *University of Illinois at Urbana-Champaign*

Previous research has considered sequential item response theory (SIRT) models for circumstances where examinees are allowed at least one opportunity to correctly answer questions. Research suggests that employing answer-until-correct assessment frameworks with partial feedback can promote student learning and improve score precision. This paper describes SIRT models for cases when test-takers are allowed a finite number of repeated attempts on items. An overview of SIRT models is provided and the Rasch SIRT is discussed as a special case. Three applications are presented using assessment data from a calculus-based probability theory course. The first application estimates a Rasch SIRT model using marginal maximum likelihood and Markov Chain Monte Carlo procedures and students with higher latent variable scores tend to have more knowledge and are better able to retrieve that knowledge in fewer attempts. The second application uses R to estimate growth curve SIRT models that account for individual differences in content knowledge and recovery/retrieval rates. The third application is a multidimensional SIRT model that estimates an attempt specific latent proficiency variable. The implications of SIRT models and answer-until-correct assessment frameworks are discussed for researchers, psychometricians, and test developers.

Bayesian vs. Bootstrap Methods for Mediation Analysis with Binary Outcomes

Paper Presentation

Charlotte Cunningham, *University of Notre Dame*
Lijuan Wang, *University of Notre Dame*

Mediation analysis has long been a popular technique in psychological research for determining the extent to which a third variable can explain the relationship between a predictor and outcome. While this technique is well understood when the mediator and outcome are normally distributed, more work is needed to determine a better method when the outcomes and/or the mediators violate this assumption. In this presentation we will focus on the case where both the mediator and outcome are binary. A simulation study conducted to compare of bootstrap and Bayesian methods for estimating and testing the indirect effect using a logistic mediation model. Results show that the performance of bootstrap methods may suffer when the sample size is small or when the outcomes and/or the mediators are rare. Explanations for the suboptimal performance and alternative implementations that may enhance performance of the bootstrap will be discussed. In contrast, the performance of Bayesian methods was more satisfactory under those conditions.

Evaluating the Performance of Cognitive Diagnosis Models Selected based on the Wald Test

Paper Presentation

Jimmy de la Torre, *Rutgers, The State University of New Jersey*
Wenchao Ma, *Rutgers, The State University of New Jersey*
Charles Iaconangelo, *Rutgers, The State University of New Jersey*

A variety of specific and general cognitive diagnosis models (CDMs) have been developed in recent years. However, selecting the most appropriate model for each item is not always apparent--on one hand, general CDMs provide better model-data fit; on the other hand, specific CDMs have more straightforward interpretations, are more stable, and can provide more accurate classifications when used correctly. But because the true model is never known, item-level model selection is a challenging process in practice. Recently, the Wald test has been proposed to determine, item by item, whether a general CDM can be replaced by specific CDMs without a significant loss in model-data fit. The current study aims to examine the practical consequence of the test by evaluating whether the attribute classification based on CDMs selected by the Wald test is better than that based on general CDMs. This study considers various sample sizes, item qualities, attribute distributions, and underlying CDMs, and finds that the CDMs selected by the Wald test have higher correct classification rates than general CDMs. This is particularly true when the sample size is small, or the item quality is low. The practical implications of this finding on the use of CDMs are discussed.

Modeling Differences in Item-position Effects in the PISA 2009 Reading Assessment Within and Between Schools

Paper Presentation

Dries Debeer, *University of Leuven*
Rianne Janssen, *University of Leuven*
Johannes Hartig, *DIPF*
Janine Bochhulz, *DIPF*

In achievement testing, it is commonly assumed that item and person characteristics are invariant with respect to administration conditions. However, it has been repeatedly shown that the position of an item within a test form affects item performance. Performance can increase ('learning/practice effect') or decrease ('fatigue effect') with item-position. Hartig and Bochhulz (2012) and Debeer and Janssen (2013) proposed an IRT model that deals with this issue by adding a 'position dimension' (i.e., persistence) to the model. Aside a general effect of item position, this model allows for individual differences in persistence, and estimates the correlation between a test-taker's ability and persistence. This paper presents a multilevel extension of the model that decomposes the variance in persistence (and ability) as well as the correlation between the two dimensions, in hierarchical data. The model is applied to data from the 2009 PISA reading assessment. Data were analyzed separately for each country. The (1) general effect of item position, (2) the individual differences in persistence within and between schools, as well as (3) the pattern of correlations between students' performance and persistence are compared across countries.

Item-Level Sensitivity to Change on the Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog)

Poster Presentation

Sien Deng, *University of Wisconsin, Madison*
Martiza Dowling, *University of Wisconsin, Madison*
Daniel Bolt, *University of Wisconsin, Madison*

We consider application of a longitudinal factor model of change (Meredith & Horn, 2001 introducing distinct latent factors related to (1) initial status differences between participants; and (2) latent change across the time points of measurement. Standard measurement invariance analyses applied to the model allow for detection of individual items that display significantly greater (or lesser) sensitivity to change as compared to their sensitivity to initial status differences. This methodology is illustrated using all available longitudinal ADAS-Cog item-level data, including the full-spectrum of disease status, from the multi-site Alzheimer's Disease Neuroimaging Initiative (ADNI I, GO, 2). A series of statistically nested models are implemented and compared. Many items demonstrate widely varying sensitivities to change as opposed to initial status.

Accounting for Measurement Error when the Dependent Variable in Multilevel Models is Latent

Paper Presentation

Ronli Diakow, *New York University*
Sophia Rabe-Hesketh, *University of California, Berkeley*

Plausible values (PVs; Mislevy, 1991) are the current standard for secondary data analysis of a latent variable. This technique relies on having a latent regression model for the construction of the plausible values that matches the intended secondary analysis. However, the current implementations of plausible values, including large-scale assessments such as PISA and NAEP, do not use a latent regression model that matches a random-effects multilevel secondary analysis. The use of such plausible values can lead to biased estimates for the variance components (Monseur & Adams, 2009; Diakow, 2010). In this paper, we propose and assess an alternative method for incorporating a latent variable as the dependent variable in a multilevel model. The proposed method uses weighted likelihood estimates (WLEs; Warm, 1989) of the latent variable, which do not rely on the specification of a conditioning model, as the dependent variable for the multilevel model. It explicitly accounts for measurement error in the WLEs by fixing part of the level-1 residual variance equal to the estimated variance of the WLEs. The performance of the proposed method is evaluated and compared to the plausible values method using simulation studies and an empirical example.

A Family of Generalized Diagnostic Classification Models for Multiple-Choice Option-Scored Assessments

Paper Presentation

Lou DiBello, *University of Illinois at Chicago*
Robert A. Henson, *University of North Carolina, Greensboro*
William F. Stout, *University of Illinois at Chicago and University of Illinois at Urbana-Champaign*

We discuss a family of restricted latent class diagnostic classification models, the GDCM-MC. It captures diagnostic information about student thinking from multiple choice assessments with response options specifically designed to link to particular latent states. Its purpose is to fully incorporate richer diagnostic information that is designed into incorrect response options. The GDCM-MC focus is on providing formatively useful information about students' thinking. Diagnostic assessments are being designed so that specific misconceptions or incomplete understandings make certain incorrect options more probable. We consider two instances of the GDCM-MC: an extended version of the Reparameterized Unified Model (alternatively, Fusion Model), denoted ERUM-MC, and an extended version of DINA, denoted EDINA-MC. The GDCM-MC addresses four DCM modeling challenges, to: (1) exploit option-based scoring, (2) include a guessing component, (3) support a latent space of 'skills' (desirable facets of thinking) as well as misconceptions/incomplete understandings, and (4) allow for a wide variety of model instantiations of differing structures (conjunctive, disjunctive, fully compensatory) and differing levels of parametric complexity that balance modeling bias with estimation variability. We present option-level-informative items and applications to real and simulated data, demonstrating effective calibration, reasonable correct classification rates, and satisfactory model-data fit.

Robust Estimation under Nonnormality and Varying Numbers of Ordered Categories

Paper Presentation

Christine DiStefano, *University of South Carolina*
Grant B. Morgan, *Baylor University*
Elizabeth Leighton, *University of South Carolina*

In recent years, many advances related to estimation techniques to accommodate coarsely categorized, non-normal data have been made. Methods include estimators that apply 'rescaling' corrections, such as Maximum Likelihood with the Satorra-Bentler correction, Diagonally Weighted Least Squares estimation, or analysis with alternative estimators, such as unweighted least squares. As robust estimators gain momentum, simulation studies are often employed to investigate the performance of the techniques. One area which has not received attention is how data from multiple scales, where scales have different numbers of scale points, may affect estimation. This situation may occur when multiple instruments are used to measure latent variables present in the same nomological network (e.g., multi-trait, multi-method research). While this practice may be encountered in practice, questions remain. Do the different numbers of scale points analyzed within the same model greatly impact results? Does one estimation technique accommodate data of this nature better than other techniques? This study will investigate situations where the number of scale points varies for data analyzed within the same model. Data will be analyzed under varying sample sizes, non-normality levels, and number of ordered categories to determine the impact on parameter estimates, standard errors, and model fit.

Measurement of Traits Obtained Via Self-Report

Keynote Address

Fritz Drasgow, *University of Illinois at Urbana-Champaign*

Many important psychological constructs are assessed by self-report measures: examples include attitudes, values, job satisfaction, personality, and vocational interests. Since the time of Likert, the vast majority of instruments measuring such traits have assumed a dominance process underlies responses. A growing body of evidence challenges the assumption of a dominance process and instead indicates that ideal point models are required for an adequate characterization of the response process. The fit of dominance and ideal point models to personality and vocational interest instruments will be summarized. The Tailored Adaptive Personality Assessment System (TAPAS), a personality assessment instrument based on an ideal point model, will described.

The Power Function of Exact Tests of the Rasch Model

Paper Presentation

Clemens Draxler, *Ludwig-Maximilians-Universität München*

In this paper, a general expression of the power function of exact tests of the Rasch model is derived. It allows the determination of the power of exact tests against various alternative hypotheses. A number of relevant examples frequently occurring in practice are discussed. With respect to computations a Monte Carlo approach is suggested enabling the approximation of the exact power in applications.

Power Analysis With Uncertainty Considered From Both Frequentist and Bayesian Frameworks

Paper Presentation

Han Du, *University of Notre Dame*
Lijuan Wang, *University of Notre Dame*

Power analysis is done often for sample size planning. In a typical power analysis, only one estimated effect size or one single value for each parameter is used in calculating power and thus it fails to consider uncertainty in the sampling distributions or in the parameter. However, in psychological studies, not only the sampling distribution of a statistic involves uncertainty but also the population effect size could be heterogeneous. Therefore, it is important to consider uncertainty and use a range of plausible values for an effect size or a parameter to conduct power analysis. In this talk, we discuss how to use information from meta-analysis to calculate power with uncertainty considered. Two power analysis methods will be introduced: the Frequentist approach based on the sampling distribution of estimated effect sizes and the Bayesian expected power based on the posterior distribution of the effect size. For each approach, we will discuss how uncertainty is included. We apply the methods to calculate power for testing Pearson's correlations and comparing mean differences with results from multiple meta-analysis studies. Power results from the two proposed methods will be compared to each other and will be compared to that from the traditional one-single-value approach.

The Evaluation of Decision Consistency Indices in Criteria-Reference Test Based on Item Response Theory

Paper Presentation

Jiaxuan Du, *Beijing Normal University*
Ping Chen, *Beijing Normal University*
Tao Xin, *Beijing Normal University*

Within the framework of Item Response Theory (IRT), several methods have been proposed to evaluate the decision consistency indices of the criteria-referenced test. This article investigated the performance of the newly-developed Lee's IRT D approach (Lee, 2010) using a simulation study and an empirical study, and the real data are collected from two assessments of a large-scale educational measurement. Results from the simulation study suggested the distribution of examinees' abilities, the number and location of the cut scores, and the score conversion rules from raw to scale could influence the decision consistency indices. To be specific, the decision consistency index decreases when using the several-for-one score conversion rule and increasing the number of cut-off score, and there are interactions between the distribution and the cut score. The real data also supported the results of the simulation study. Thus, we conclude Lee's IRT D method is reliable for evaluating decision consistency indices. Suggestions are made that researchers should clearly describe the situation of test and report the decision consistency indices under each situation for individual examinees across the range of possible scores. It will be useful to apply Lee's IRT D method to the field of large-scale educational measurement.

Some Thoughts on Validity and Model Fit

Invited Speaker

Michael Edwards, *The Ohio State University*

Although there are many definitions offered for validity (and validation), my favorite is still Messick's 1989 (p.13, emphasis in original), "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment." While recently reading the (very excellent) Frontiers of Test Validity Theory (Markus & Borsboom, 2013) it struck me that just as scores from a test require validity evidence to support their use, so to do various model fit indexes used in psychometrics. In this talk, I will review some similarities in the problem of validity found in testing and the evaluation of model fit. I will highlight some progress that we have made as well as some areas I believe more work is needed to be done. I conclude with some general conclusions and suggested future directions.

Diagnosing Broad and Narrow Skills with the Multicomponent Latent Trait Model: A Study of Middle School Mathematics

Paper Presentation

Susan E. Embretson, *Georgia Institute of Technology*

The multicomponent latent trait model for diagnosis (MLTM-D; Embretson & Yang, 2013) is a conjunctive item response model that is hierarchically organized to include broad and narrow skills. A two-stage adaptive testing procedure was applied to diagnosis skill mastery in middle school mathematics and then analyzed with MLTM-D. The current study presents results on both broad and narrow skill mastery. Reliability of skill possession is assessed using an uncertainty analysis with multiple imputation.

A Generalized Partial Credit FACETS Model for Investigating Order Effects in Self-Report Personality Data

Paper Presentation
Susan Embretson, *Georgia Tech*
Heather Hayes, *Georgia Tech*

Despite its convenience, the process of self-report in personality testing can be impacted by many biases. One bias is the order effect in which characteristics of an item response are impacted not only by the content of the current item but also the accumulated exposure to previous, similar-content items. Previous investigations of this effect have been rooted in classical test theory and have found that item reliabilities increase with the item's serial position in the test. The purpose of the current study was to more rigorously examine order effects via item response theory. A new model, the Generalized Partial Credit FACETS model (GPCFM), was constructed to quantify the order effect. Position of an item serves as a new type of facet that contributes to the item response, not only via its impact on an item's location parameter, but also its discrimination. Thus, the GPCFM differs from previous generalizations of the FACETS model in that the item discrimination parameter is modified to include a serial position effect. Simulation analyses supported viability of the model, and the results of real data analysis supported the GPCFM as having superior fit relative to competing models while identifying order effects across multiple personality traits.

A Flexible and Efficiently Estimated High-dimensional Response Style Model

Paper Presentation
Carl F. Falk, *University of California, Los Angeles*
Li Cai, *CRESST / University of California, Los Angeles*

Despite recent advances in the modeling of response styles, the ability to estimate full-information models with a high number of correlated substantive factors and/or several different continuous response style factors is sometimes limited by the estimation approach or the type of response style model that is chosen. In this talk, we present an item response theory model that is able to model multiple user-defined response styles such as extreme response style, midpoint response style, and so on, across multiple constructs of interest. The model is based on a new parameterization of the multidimensional nominal response model that separates estimation of overall item slopes from the scoring functions (or order of categories) for a particular item and latent trait. This feature allows for item slopes for response style and substantive factors to vary freely across items. The model uses an efficient estimation algorithm (Metropolis-Hastings Robbins-Monro) that allows for a high number of correlated latent constructs and recursive estimates of parameter standard errors. The model is demonstrated using examples from the smoking initiative of NIH's Patient Reported Outcomes Measurement Information System and is tested in a simulation study.

Moderated Mediation with MIMIC Models: SEM Analysis of Personal Resources, Perceived Age-related loss, and Aging Satisfaction

Paper Presentation

Yuyu Fan, *Fordham University*
Ying Liu, *Fordham University*
Daniela S. Jopp, *Fordham University*

Despite the usefulness of Multiple Indicators Multiple Causes (MIMIC) models, there has been little application of them in aging studies. In this paper, we first justify the MIMIC model for personal resources using the tetrad test, and then apply this model to examine the mediation relationship between personal resources, perceived age-related loss and aging satisfaction with perceived age-related loss as the mediator and aging satisfaction as the outcome variable. Moreover, the effect of two binary moderators on the mediation effect, gender and age, are examined respectively. Results show that the MIMIC model is more appropriate than the common factor model for personal resources both substantively and statistically. The direct effect of personal resources on aging satisfaction accounts nearly two thirds of the total effect, and the indirect effect accounts over one third of the total effect. In addition, there are significant differences between males and females, young adults and old adults on the mediation effect. Those findings support the usage of MIMIC models and call for our attention on the gender and age differences in aging study.

An Effect Size Measure for Partial Factor Invariance

Poster Presentation

Holmes Finch, *Ball State University*
Brian French, *Washington State University*

Factor invariance (FI) is a central concept in psychological measurement. When latent variable model parameters (e.g. factor loadings, intercepts) lack invariance across groups, scores on the scale may not have equivalent meaning for individuals from different groups, thereby compromising the scale's score validity (Millsap, 2011). While FI is desirable for psychological scales, it is a rare commodity in practice (Millsap & Meredith, 2007). Thus, comparisons are made under partial invariance (PI), where only some model parameters are invariant (Byrne, Shavelson, & Muthén, 1989). While methods for detecting PI are reasonably well developed, quantifying the magnitude and influence of PI on scores is less developed. The goal of this study is to introduce a new effect size measure to quantify the impact of PI with respect to factor loadings and intercepts on scale score means and their comparison across groups, and to assess the new measure's performance. The manipulated factors in the simulation study include sample size, group size ratio, number of factor indicators, proportion of invariant factor indicators, and magnitude of group latent mean differences. Outcomes to be reported are group mean estimation accuracy and accuracy of group mean comparisons.

Rule-based Methods for Clinical Prediction Problems

Paper Presentation

Marjolein Fokkema, *Vrije Universiteit*
Niels Smits, *Vrije Universiteit*
Henk Kelderman, *Vrije Universiteit*

Meta-analyses comparing the accuracy of clinical versus actuarial prediction methods have shown actuarial methods to outperform clinical methods, on average. However, actuarial methods are still not widely used in clinical practice, and there has been a call for the development and explanation of actuarial prediction methods for clinical practice. In this paper, we argue that rule-based methods may be more appropriate than the linear main effect models that are usually employed in prediction studies. Rule-based methods may be more appropriate from a data analytic, decision analytic, as well as a substantive perspective. In addition, decision rules can be represented as fast and frugal trees (FFTs), which have gained popularity in the area of decision analysis, due to their usability. We provide an illustration of the application of the RuleFit algorithm, by reanalysis of a data set on prediction of the two-year course of depression and anxiety disorders. The decision rules derived by RuleFit are compared in terms of sparsity and accuracy, with the results of logistic regression analysis, originally applied to the dataset.

Higher-level Response Modeling using Plausible Values

Paper Presentation

Jean-Paul Fox, *University of Twente*

A striking feature of large-scale surveys is that individual achievements are explicitly measured and modeled, although inferences are required at higher levels. Large-scale standardized tests such as PISA (Programme for International Student Assessment) have the objective to evaluate education systems, achievements of subgroups and countries. In contrast to this, individual-level parameters are not of primary interest but serve to construct higher-level inferences and to model dependencies between response observations. From this perspective, a marginalized lower level modeling approach will be proposed partly to avoid the extreme computational challenges associated with the prevalent heavy lower-level parameterization to model disaggregate data (Camilli, Fox, Kim, 2014). Using the underlying latent-response formulation, the distribution of the errors can be modeled to account for the nesting of subjects and the heterogeneity among subjects. Using recent developments in sampling techniques and rank likelihood methods, clustered latent response data are sampled (using only the ranks of the observed discrete data). Within this modeling framework, it will be shown how plausible values can be obtained and used to support higher-level modeling, without assuming local independence at the level of subjects. Extensions will be proposed to impute multivariate higher-level latent variables at higher levels of observations.

Maximum Likelihood Item Easiness Models for Test Theory Without An Answer Key

Paper Presentation

Stephen L. France, *University of Wisconsin-Milwaukee*
William H. Batchelder, *University of California – Irvine*

Cultural Consensus Theory (CCT) is a data aggregation technique with many applications in the social and behavioral sciences. We describe the intuition and theory behind a set of CCT models for continuous type data using maximum likelihood inference methodology. We describe how additive and multiplicative bias parameters can be incorporated into these models. We introduce two extensions to the basic model in order to account for item rating easiness/difficulty. The first extension is a multiplicative model and the second is an additive model. We show how the multiplicative model is related to the Rasch model. We describe several maximum-likelihood estimation procedures for the models and discuss issues of model fit and identifiability. We describe how the CCT models could be used to give alternative consensus based measures of reliability. We demonstrate the utility of both the basic and extended models on a set of essay rating data. We describe a MATLAB based software package that implements the techniques described in the presentation.

Extensions to Item Mapping Graphs

Paper Presentation

Rebecca Freund, *University of California, Berkeley*
David Torres Irribarra, *University of California, Berkeley*

Graphical representations of results are an important tool in interpreting and exploring analyses. In item response theory, one particularly useful way of representing results is the item map, which displays item parameters and person proficiencies on a common scale. To link items to persons, these maps commonly represent item locations as the ability level at which respondents' probability of responding correctly is predicted to be 0.5. This paper presents some extensions to item maps that can further improve their usefulness to researchers and end users. One such extension that we present is the use of a dynamic graph that allows users to observe how the location and order of items in the item map can change if we represent item locations as the ability level at which respondents' probability of correct response is equal to some other value (besides 0.5). Another is the implementation of individual item maps, which represent item locations with reference to a single respondent, can display expected response patterns as a function of individual ability estimates. These new graphical methods are implemented using R, a free and widely used statistical programming software. They can accept output from other statistical packages as data to generate the displays.

Multiple Dependent Dirichlet Process Rating Model

Paper Presentation

Ken A. Fujimoto, *University of Illinois at Chicago*
George Karabatsos, *University of Illinois at Chicago*

Item response theory (IRT) models are often used to analyze data arising from rating scale questionnaires. Commonly used IRT rating scale models assume that the rating category threshold parameters are the same over examinees. This assumption can easily be violated when there is differential item functioning (DIF) due to known or latent subgroups of examinees in the data sample. To address this practical psychometric problem, we introduce a novel, Bayesian nonparametric IRT model for rating scale analysis. The model is an infinite-mixture of Rasch partial credit models, with the mixture based on a multiple Dirichlet process (mDP). The model treats the category thresholds as the random parameters that are subject to the mixture, which has (stick-breaking) mixture weights and atoms that are covariate-dependent. That is, this model allows the distribution of the category thresholds to vary as a function of covariates, which for this presentation are item indicators. Thus the mixing distribution can inform whether an item has DIF across latent subgroups of examinees in a sample. We illustrate the effectiveness of this new model through the analysis of real and simulated data.

Modeling Strategies

Paper Presentation

April Galyardt, *University of Georgia*
Brian Junker, *Carnegie Mellon University*

Strategy choice is one important dimension that distinguishes expert performance from that of novice. Even when a novice is able to successfully complete a task, their strategy is often much less efficient than that of an expert. While there have been a few attempts at modeling student strategy usage, the area has remained largely unexplored for several reasons. First, is that all students switch strategies between tasks, and often within tasks. Analysis at this fine grained level has been difficult, if not impossible, when data the only includes information on correct/incorrect. However, fine grained data collection, such as from an intelligent tutoring system has become more common, making strategy modeling more feasible. We examine the psychological literature on strategies and the features of student performance that can be used to distinguish strategies in different contexts. We compare existing psychometric models in their ability to analyze these necessary dimensions of student performance, and suggest directions for future work to address this need.

Relations between Aggression and Social Status in Chinese Students- Moderating Effects of Academic Achievement

Poster Presentation

Yiran Gao, *Beijing Normal University*
Yufang Bian, *Beijing Normal University*

We examined the moderating effects of academic achievement on relations between aggression and social status in Chinese children. The result shows that the academic achievement has significant moderating effects on relations between physical aggression and peer rejection, social standing, peer influence. And it also has significant moderating effects on relations between relational aggression and peer rejection, loneliness, and peer influence.

Latent State-Trait Models for a Combination of Random and Fixed Situations

Paper Presentation

Christian Geiser, *Utah State University*
Kaylee Litson, *Utah State University*
Jacob Bishop, *Utah State University*
Brian T. Keller, *Arizona State University*
G. Leonard Burns, *Washington State University*

Latent state-trait (LST) models (Steyer, Ferring, & Schmitt, 1992) allow separating person-specific (trait) effects from (1) effects of the situation and person × situation interactions, and (2) random measurement error in purely observational studies. Typical LST applications use measurement designs in which all situations are sampled randomly and do not have to be known for any individual. Limitations of conventional LST models for only random situations are that traits are implicitly assumed to generalize perfectly across situations, and that main effects of situations are inseparable from person × situation interaction effects because both are measured by the same latent variable. In this article, we show how these limitations can be overcome by using measurement designs in which two or more random situations are nested within two or more fixed situations that are known for each individual. We present extended LST models for the combination of random and fixed situations and show that the extensions allow (1) conceptualizing traits as situation-specific and (2) isolating person × situation interactions from situation main effects. We demonstrate that the new modeling approach can be applied with both homogenous and heterogeneous indicators in either the single- or multi-level structural equation modeling frameworks. Advantages and limitations of the new models as well as their relation to other approaches for studying person × situation interactions are discussed.

Evaluation of the Reliability of Composite Scores using IRT

Paper Presentation

Cees Glas, *University of Twente*
Lisette Siemaons, *University of Twente*

Composite scores from multiple variables are constructed to form reliable and valid measures of constructs. A method is proposed to determine weights that optimize the reliability of a composite score in the framework of IRT. Two conceptualizations of reliability in IRT are considered. These reliability indices are then combined with two often-used estimation methods for IRT: marginal maximum likelihood and the Bayesian MCMC framework. Further, methods for obtaining the confidence or credibility regions for these weights are presented.   The empirical example motivating this research is the 28-joint Disease Activity Score (DAS28), the most widely used disease activity index in rheumatoid arthritis (RA). The index combines scores on a 28-tender joint count (TJC28), a 28-swollen joint count (SJC28), a patient-reported visual analog scale for general health (GH), and a non-specific acute phase reactant of systemic inflammation (either the erythrocyte sedimentation rate [ESR] or the C-reactive protein [CRP]) into a composite measure of disease activity. The data are combined into a multidimensional latent variable model, and the optimal weights are determined for the two conceptualizations of global reliability in IRT and the two estimation frameworks.

A Novel Approach to Evaluate Item Pools: The Item Pool Utilization Index

Paper Presentation

Emre Gonulates, *Michigan State University*
Mark Reckase, *Michigan State University*

A novel approach to evaluate item pools will be proposed. An index to quantify the quality of an item pool for a given adaptive test will be presented. This index can be used to compare different item pools or diagnose the deficiencies of a given item pool by quantifying the amount of deviation from a perfect item pool. We will conduct a study to: (1) show how to compute this index for a given item pool, (2) compare different item pools using this index, (3) explore and show the added value of this index compared to existing ways to evaluate the quality of the Computerized Adaptive Tests (CAT). Simulations will be performed to show the usefulness of this index. Initially, adaptive tests with different specifications will be simulated to exhibit the performance of this index. Item pools will be compared using item pool utilization index and various other CAT outcomes. Test developers can use this index as a diagnostic tool to evaluate the sufficiency of their item pool. This index will answer the question: Does the item pool provide examinees appropriate items, or does it fall short of this requirement?

On the Distribution of Test Scores Conditional on Ability and Some Related Equating Methods Based on it

Paper Presentation

Jorge Gonzalez, *Pontificia Universidad Católica de Chile*
Marie Wiberg, *USBE, Umeå University*
Alina A. von Davier, *Educational Testing Service*

Lord (1980) pointed out that using IRT, the frequency distribution of test scores conditional on a given ability, $\varphi(x|\theta)$, can be obtained. For the case when x is the summation of the correct responses, Lord recognized that an explicit form of $\varphi(x|\theta)$ was difficult to obtain in a simple form. However, Lord wrote the expectation of this distribution. Lord and Novick (1968) defined an individual's "true score" as the expected value of the individual's observed score. Using the expectation of the conditional distribution and this classical test theory definition, Lord developed the IRT true-score equating. What would have happened if not only the mean but the complete conditional distribution was derived analytically? In this talk, an explicit mathematical form of the conditional distribution $\varphi(x|\theta)$ will be presented. We will show that the mean and variance of this distribution coincides with the ones derived by Lord only using standard arguments. We will also show that the well known recursive algorithm presented by Lord and Wingersky (1984) corresponds to an identity derived using the moment generating function of x . A discussion about the potential applicability of this distribution in local equating (LE) and the relationship between IRT true-score equating and LE will be also presented.

The Effect of Different Instructions on Examinee Behavior in a Speeded Computer Adaptive Test

Paper Presentation

Darrin Grelle, *CEB*
Sara Gutierrez, *CEB*

Past literature on how test instructions influence examinee behavior has generally focused on correction for guessing with the majority of research published in the 1970's and 1980's prior to computer based testing/IRT scoring becoming the norm. The proposed policy capturing study will evaluate the effect of different instructions on examinee behavior taking a computer adaptive deductive reasoning test used for employee selection. A scoring adjustment is implemented for examinees that do not complete, and examinees cannot skip questions. This study will examine the effects of different instructions on final test scores, number of questions answered, and question response times. Hierarchical linear modeling will be used to evaluate level two effects of race, gender, age and personality characteristics as research has shown that these variables affected how examinees responded to different instructions. This study will determine if these relationships hold in a semi-speeded, computer adaptive context. Three hundred examinees will take one of the three versions of the test (differing only in instructional content) once a week for three weeks. The instructions will either (1) not mention the penalty, (2) indicate that examinees must answer at least 2/3 of the items, or (3) inform examinees of the penalty for not completing.

A State Space Approach to Canonical Correlation Analysis

Paper Presentation

Fei Gu, *McGill University*

Canonical correlation analysis (CCA) is a generalization of multiple correlation for analyzing the relationship between two sets of variables. Bagozzi, Fornell, and Larcker (1981) showed that CCA could be subsumed by the structural equation modeling (SEM) framework as a special case, in which two MIMIC models need to be estimated for a particular pair of canonical variates. The SEM approach to CCA has two major advantages over conventional CCA: 1) providing the asymptotic standard error for individual parameter, and 2) testing statistical significance of a canonical correlation using likelihood ratio test between nested models. However, the SEM approach is less efficient regarding model specification and parameter estimation because two MIMIC models have to be fit separately for each pair of canonical variates. In this paper, we propose a state space approach that is more efficient than the SEM approach regarding model specification and parameter estimation, while maintaining the two existing advantages. Specifically, only one state space model is estimated for each pair of canonical variates by the state space approach, with fewer parameters to be estimated. Moreover, the proposed state space approach provides additional advantages that may extend the conventional CCA to capture dynamics.

Power Computation for Likelihood Ratio Test in Latent Markov Models

Paper Presentation

Dereje W.Gudicha, *Tilburg University*
Jeroen K. Vermunt, *Tilburg University*
Verena D. Schmittmann, *Tilburg University*

This paper presents power computation methods for hypotheses that can be specified on the parameters of latent Markov (LM) models. The parameters likely to be of most interest for LM models are the transition probabilities, the probabilities for switching between the latent states of successive measurement occasions. For these parameters, hypotheses are formulated and their test statistics, namely the likelihood ratio (LR) is discussed. For some of these hypotheses standard asymptotic chi-squared tests can be used, implying that also the power computation can be based on asymptotic distributions. In other situations we have to rely on bootstrap methods. Design factors affecting the power of this test are also studied. A numerical study conducted to illustrate the proposed power computation procedures showed that in addition to the usual design factors (e.g., effect size, sample size, and Type I error), power computation for LM models should also involve a number of other design factors -- namely, the number of time points, the number of indicator variables, the strength of association between states and indicator variables, the number of states, the size of initial states, and the structure of state transition probabilities.

A Forgotten Rasch Model: A New Approach to Dimensionality and Distractor Analysis of Multicategorical Data

Paper Presentation

Can Guerer, *LMU Munich*
Clemens Draxler, *LMU Munich*

This study focuses on the application of the multidimensional multicategorical Rasch Model as suggested by Rasch (1961) for assessing properties of dimensionality and distractors within the framework of cML estimation. The multidimensional multicategorical Rasch Model offers a well suited general model, which comprises a manifold of widely used models that can easily be derived from it by applying appropriate restrictions, as e.g., the binary Rasch-Model and the Rating Scale Model. Additionally an alternative parameterization can be provided, which interprets the model as a unidimensional model with different discrimination parameters per category. By restricting the parameter space appropriately one can then derive tests based on cML theory for dimensionality as described by Rasch (1961) and Andersen (1972) and hence also for distractor analysis. The theoretical properties of the model will be presented and discussed, leading to an application to real data.

Doubly Robust Estimation of Treatment Effects from Observational Multilevel Data

Paper Presentation

Courtney Hall, *University of Wisconsin-Madison*
Jee-Seon Kim, *University of Wisconsin-Madison*

When randomized experiments cannot be conducted in practice, propensity score (PS) matching and regression techniques are frequently used for estimating causal treatment effects from observational data. These methods remove bias caused by baseline differences in the treatment and control groups via the selection mechanism and the outcome mechanism, respectively. Instead of using a PS technique or an outcome regression singly, one might use a doubly robust estimator which combines a PS technique (matching, stratification, or inverse PS weighting) with an outcome regression to attempt to address the confounding issue via both the selection and outcome mechanism simultaneously. Theoretically, if one or both of the mechanisms are correctly specified with regard to functional form and error structure, a doubly robust estimator will produce an unbiased average treatment effect (ATE). Doubly robust estimators are not yet well studied for multilevel data where selection into treatment takes place among level-one units within clusters. Using a simulation study, we show how well a doubly robust estimator which combines PS stratification with regression removes selection bias when estimating the ATE from two multilevel populations compared to the bias that is removed by PS estimators and regression estimators alone.

Applications of Classification Consistency in Latent Class Models

Paper Presentation

Peter Halpin, *New York University*

If a person is classified into the same category based on each of multiple administrations of a measurement instrument, the classification is said to be consistent. Much existing work has sought to estimate classification consistency in the absence of multiple administrations, and based on psychometric models in which the underlying construct is continuous. In this paper I consider approaches to and uses of classification consistency in application to latent class models. I focus on cases where multiple administrations are available over various facets, to set up a G-study situation. I consider a cross-validation approach to classification consistency to (a) select the number of latent classes, and (b) select the number and type of administrations required to calibrate the model. The approach is illustrated with application to several instruments used for in-classroom observation of teaching.

Determinants of Achievement Level of Korean Students in TIMSS 2011

Poster Presentation

Jung-A Han, *Korea Institute for Curriculum and Evaluation & Yonsei University*
Seungeun Chung, *Yonsei University*
Sang-Jin Kang, *Yonsei University*

The purpose of study is to ascertain determinants of mathematics and science achievement levels of Korean students in TIMSS 2011. To do this, hierarchical generalized linear model methods were used.

Diagnostic Classification Modeling with Multiple Higher-order Dimensions

Paper Presentation

Mark Hansen, *UCLA*

When fitting diagnostic classification models, the number of possible attribute profiles increases exponentially with the number of attributes. To facilitate estimation, it is thus often desirable to specify a higher-order structure, as in the higher-order DINA model proposed by de la Torre and Douglas (2004). Such a model may be appropriate when attribute profile probabilities can be adequately explained by one continuous, higher-order dimension. However, there are situations in which we might expect that a single dimension will be inadequate. When a test is administered multiple times, for example, a model with correlated higher-order dimensions might be appropriate. However, a multidimensional higher-order structure may sometimes be needed for data from a single test administration. In tests of language proficiency, for example, items may align to proficiency standards, and standards in turn may be grouped within modalities of language (e.g., receptive, productive, interactive). In this study, we examine the utility of diagnostic classification models with multidimensional higher-order structure. Through simulation, we examine parameter recovery and explore the effects of model misspecification, describing the extent to which fitting a model with one higher-order dimension affects accuracy of classification with respect to attributes. Two empirical illustrations are then presented.

Assessing Students' Performances from Process Data in Scenario-based Tasks: An Edit Distance Approach

Paper Presentation

Jiangang Hao, *Educational Testing Service*
Zhan Shu, *Educational Testing Service*
Yoav Bergner, *Educational Testing Service*
Mengxiao Zhu, *Educational Testing Service*
Alina von Davier, *Educational Testing Service*

Students' activities in scenario-based tasks (SBT) are generally stored in the corresponding log files, and can be characterized by a sequence of time-stamped actions of different types. For a subset of the SBTs where the temporal variation is not a major concern, such as the WELL task in the NAEP TEL, the process data can be well characterized as a string of characters (action string, hereafter) if we encode each action name as a single character. Therefore, comparing students' performances is equivalent to comparing these character strings.   In this paper, we report our work on evaluating students' performances by comparing how far their action strings are from the action string that corresponds to the best performance, where the farness or closeness is quantified by an edit distance between the strings. Specifically, in this work, we choose the Levenshtein distance, which is defined as the minimum number of insertion, deletion and replacement needed to convert one character string to another. Our results show a strong correlation between the edit distances and the scores obtained from the scoring rubrics of the WELL task, implying the edit distance approach is a potentially valid and efficient way to characterize process data.

The Impact of Random Differential Item Functioning (DIF) when No Fixed Effect of DIF Exists

Paper Presentation

Jared K. Harpole, *University of Kansas*
Carol M. Woods, *University of Kansas*

Educational and psychological data are often collected by sampling participants in a number of different higher level units. Sampling in this way creates a 3 level hierarchical structure which, if not taken into account, may cause problems with statistical inference (Snijders & Bosker, 2012). Differential item functioning (DIF) occurs when item properties differ for members of different groups who are matched on levels of the latent variable. The purpose of the present study is to assess the impact on testing for DIF using a three-level Rasch model when the population value of fixed DIF is zero, but has a random effect across level-three (L3) units. A simulation study will be presented in which Type I error and parameter bias are evaluated for models with population DIF effects with a zero mean but random variability across L3 units. A 5-way factorial design with 32 conditions [2 (model L3 units) x 2 (random DIF) x 2 (L3 latent ability) x 2 (level 2 units) x 2 (L3 units)] will be conducted. Results and conclusions will be discussed.

Monte Carlo Studies in Quantitative Research in Education and Psychology

Paper Presentation

Michael Harwell, *University of Minnesota*
Nidhi Kohli, *University of Minnesota*

Monte Carlo studies represent an important tool for investigating the behavior of statistical procedures. Yet the potential of these studies to inform statistical practice has not been fully realized in part because recommendations to treat these studies as statistical sampling experiments (e.g., Hoaglin & Andrews, 1975; Spence, 1983) have not been widely adopted. We synthesize and extend existing literature with four complementary recommendations for these studies: (i) A clear rationale for why a Monte Carlo study is needed and how the study adds to the literature should be provided, (ii) The conceptualization and execution of these studies should draw heavily on established principles of research design (e.g., mixed effects ANOVA models, enhancing external validity), data analysis (e.g., response surface and meta-analytic procedures), and reporting of results, (iii) Evidence of the accuracy of the simulated data should be provided, (iv) The description of a Monte Carlo study should allow readers to evaluate what was done, why, and to assess a study's contribution to statistical practice and to a research program. The goal is to encourage methodological researchers to exploit the strengths of Monte Carlo studies in ways that inform statistical practice. Our recommendations are illustrated with simulated data.

Analyzing Process Data in Problem-solving Items with n-gram Model: Insights from a Computer-based Large-scale International Assessment

Paper Presentation

Qiwei He, *Educational Testing Service*
Matthias von Davier, *Educational Testing Service*

In computer-based assessments, the process data in log files provide new insights into behavioral processes of task completion. The present study analyzes process data collected from a complex task that is part of the scale of problem-solving in technology-rich environment (PS-TRE) in a computer-based large-scale international program PIAAC. The purpose of this study is twofold: first, to extract and identify robust indicators of sequential actions that lead to success or failure in a PS-TRE item; and secondly, to compare the extracted sequence patterns across selected PIAAC countries. Motivated by the methodologies in natural language processing and text mining, we used the n-gram representation and a chi-square feature selection model in analyzing the process data that have a similar structure as written language. Initial results show that the test takers who respond correctly were more likely to use tool functions such as searching or sorting to structure the problem whereas respondents who produce an incorrect response were more likely to show hesitative behaviors such as repeatedly clicking cancel buttons and using the 'Help' function. Comparisons across the selected countries regarding the problem-solving process are also discussed in the paper.

A Method for Simulating Univariate and Multivariate Non-normal Distribution through the Method of Percentiles

Paper Presentation

Todd Christopher Headrick, *Southern Illinois University Carbondale*

Power method polynomial transformations are commonly used for simulating continuous non-normal distributions with specified moments. However, conventional moment-based indices such as skew and kurtosis are often unavailable or not defined. Some examples would include (i) standardized tests-score reports or (ii) certain Student t-distributions. It is more often the case that percentiles are reported or defined. In view of this, we derive a power method polynomial transformation based on the method of percentiles (MOP). The derivation produces a convenient system of equations that yields solutions to coefficients that are available in closed-form. As such, the solutions are unique whenever they exist and thus obviate the need for non-linear equation solving. The primary focus of the derivation is on polynomial transformations of order five. Comparisons between the proposed MOP procedure are made with its conventional method of moments (MOM) counterpart i.e. Fleishman (1978) or Headrick (2002) polynomials. The proposed MOP methodology is also extended for the purpose of simulating multivariate non-normal distributions with specified Spearman correlations. Monte Carlo results demonstrate that estimators based on the proposed MOP methodology are superior or comparable to their corresponding conventional MOM estimators (e.g., skew, kurtosis, Pearson correlation) in terms of relative bias and relative standard error.

Regularization with the Coefficient of Variation in Loss Functions for Nonmetric Multidimensional Unfolding

Paper Presentation

Willem J. Heiser, *Leiden University*
Frank Busing, *Leiden University*

The usual Stress loss functions for Nonmetric Multidimensional Unfolding need regularization in order to avoid undesirable solutions with low stress value that are uninformative (degeneracies). It was shown earlier that regularization on the variance is not sufficient, and that regularization on the variation coefficient is better. We now discuss a formulation is that is simpler than the earlier one, and seems to get better results in the test data sets used so far. In the row-conditional case, it gives relatively smaller weight to individuals who do not fit well, so that their influence on the total solution is diminished.

Latent Class Detection in the Case of Qualitative Group Differences: A Comparison of Taxometrics and IRT Mixture Modeling

Paper Presentation

Robert Hillen, *Tilburg University*
Wilco Emons, *Tilburg University*
Jelte Wicherts, *Tilburg University*
Klaas Sijtsma, *Tilburg University*

Taxometrics is a popular statistical procedure in the fields of psychopathology and psychiatry for detecting latent classes or taxonicity, as it is referred to in the field of taxometrics. The method is easily applicable and does not rely on distributional assumptions. The performance of taxometrics in detecting latent classes has been frequently studied under varied data conditions, but rarely under conditions that are characteristic of psychological data. Such conditions include low levels of class separation and qualitative groups differences defined as non-invariant measurement models across groups. Mixture models are less frequently used but offer an alternative to taxometrics as these models allow for specific hypothesis testing of covariance structures and can account for qualitative group differences. The goal of this study is to investigate the performance of different taxometric procedures in detecting latent classes under these conditions. We also compare the performance of these taxometric procedures to IRT mixture modeling in detecting latent classes. The results provide insight into the power of taxometrics and mixture modeling in detecting latent classes in the fields of psychopathology and psychiatry.

A Simulation Study: The Precision and Efficiency of Computer Adaptive Tests

Poster Presentation

Tiffany Hogan, *Georgia State University*
Chris T. Oshima, *Georgia State University*

The purpose of this paper was to demonstrate the efficiency and precision of computer adaptive testing. This paper used real data to compare the results of a fixed and variable-length computer adaptive simulation to a full length. The Computer Adaptive Test Simulation (CATSim) Program was used to measure the precision of computer adaptive tests (CAT) by comparing the estimated CAT theta values from each simulation with the true thetas values obtained from the 586 examinees on a full length test. Simulation results were compared using standard error and classification termination points. The results of both simulations revealed that the Variable Length (VL)-CAT and Computerized Classification Test (CCT) had a larger the mean standard error than the full length test. The relationship between full length test and CAT thetas were strong and positively correlated in both VL-CAT and CCT. The full length test had a lower standard error than the VL-CAT and CCT because of the greater number of items. Although the full length test had a smaller standard error, the difference was minor when compared to the reduced amount of time (efficiency) in administering a computer adaptive test. This makes the computer adaptive test more efficient without compromising precision.

A Comparison of Commonly Used DIF Techniques in an MST Setting

Paper Presentation

Likun Hou, *American Institute of Certified Public Accountants*
Jonathan Rubright, *American Institute of Certified Public Accountants*
Oliver Y. Zhang, *American Institute of Certified Public Accountants*

Although recent research on differential item functioning has not generated much enthusiasm due to the historical attention to DIF in psychometrics (Wainer, 2010), this work remains an important aspect of test fairness and validity (Kane, 2013). While DIF has been extensively studied for linear tests, the literature is fairly narrow for computer adaptive tests and virtually nonexistent in multi-stage adaptive testing (MST; Gierl, Lai, & Li, 2013). This study is the first to systematically examine the performance of commonly used DIF procedures [Mantel-Haenszel, CATSIB, logistic regression (LR) fit and LR coefficient approaches] in the context of different group impacts, sample sizes, DIF sizes, proportions of DIF items, and extent of operational item contamination, with data simulated in an MST setting. Procedures are evaluated via Type 1 error and power rates, with ANOVAs to identify significant differences. Accuracy of effect size categorization, used operationally to make item-level decisions, is also evaluated. Results show CATSIB had Type-1 error rates closer to the nominal level. All five procedures had excellent power with larger samples and large DIF sizes. Significant differences were found in how procedures categorized DIF items, with CATSIB the most accurate in correctly categorizing moderate and large DIF items.

A New Estimation Procedure for Partial Least Squares Path Modeling

Paper Presentation

Heungsun Hwang, *McGill University*
Yoshio Takane, *McGill University*
Arthur Tenenhaus, *Sup,lec*

We propose a new procedure for estimating component weights in partial least squares path modeling. The proposed procedure aims to minimize a single least-squares criterion that includes Mode A and Mode B as special cases. An alternating least-squares algorithm is developed to minimize the criterion consistently. This procedure can be viewed as an alternative to the path weighting scheme, in which component weights are estimated optimally in a least-squares sense, while simultaneously considering both magnitudes and directions of the relationships between latent variables. We show that the proposed and existing procedures result in quite similar or identical solutions when they are applied to real and simulated data.

Unfolding Item Response Model Using Best-Worst Scaling: Measurement of Attitude Toward Tardiness

Paper Presentation

Kazuya Ikehara, *Waseda University*

In attitude measurement and sensory test, we usually use an unfolding model in which response probability is formulated by distance between locations of the person and the stimulus. In this paper, we proposed an unfolding item response model using best-worst scaling (BWU model) in which person chose the best and the worst from repeatedly presented subsets of stimuli. We also formulated an unfolding model using best scaling (BU model), and compared the accuracy of estimates between the BU model and the BWU model. Simulation study showed that the BWU model performed much better than the BU model in terms of bias and root mean square errors of estimates. With reference to the research of Usami (2011), proposed models were applied to actual data to measure the attitude toward tardiness. As a result, we found that the estimates of stimuli between proposed models and Usami's were very similar.

Item Response Model for Partially Ordered Data and Application to Common Sense Beliefs about Disease

Paper Presentation

Edward Ip, *Wake Forest School of Medicine*

There is a general lack of analytic methods for data structures other than ordered or nominal categories. Particularly, partially ordered set (poset) responses, for example, responses that contain categories from strongly agree to strongly disagree, as well as a don't know category- which are quite commonly encountered in psychological and other social sciences, were often not appropriately handled. Poset responses were routinely collapsed and then analyzed as either rank ordered or nominal data, leading to the loss of nuanced information that might be present within poset categories. This paper introduces a general class of item response models for multiple response data that contain poset structures. The inferential object of interest is the latent trait that purportedly drives covariation of poset response patterns. It is proved that a weakly ordered chain can be systematically delineated from any poset structure, connected or otherwise, and a simple coding algorithm can then be applied to poset response data such that standard software could be directly used for the purpose of item estimation and individual scoring. Simulation experiments and a real data set regarding beliefs about diabetes illustrate the poset item response model and the related methods.

A Pragmatic Perspective on Measurement

Paper Presentation

David Torres Irribarra, *University of California, Berkeley*

Throughout history the technical understanding of measurement has been influenced by key philosophical perspectives such as the "classical view" of measurement (Michell, 2005), operationalism (Bridgman, 1927) and the representational theory of measurement (Pfanzagl, 1968; Krantz, Luca, Suppes & Tversky, 1971). Despite their differences, these traditions share more or less explicitly the core assumption that the role of measurement is to adequately mirror reality. Pragmatism breaks with this assumption, replacing it with the notion that the role of language, knowledge, science and measurement is to help us cope with reality (James, 1907; Rorty, 1999). From a pragmatic perspective, measurement is another instrument in our toolbox, which we use to interact with the world. I examine the implications of adopting a pragmatic reading of measurement regarding central definitional aspects such as (a) what is measurement, (b) what can be measured, and finally, (c) how a pragmatic standpoint would affect the notion of ability as usually discussed in psychometrics.

Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment

Paper Presentation

Ruslan Jabrayilov, *Tilburg University*
Wilco Emons, *Tilburg University*
Klaas Sijtsma, *Tilburg University*

Clinical psychologists are advised to assess clinical and statistical significance of change when assessing change in individual patients (Jacobson & Truax,1991). Individual change assessment can be conducted using either the methodologies of classical test theory (CTT) or item response theory (IRT). There is a growing optimism among researchers regarding the possible advantages of using IRT over CTT when constructing and scoring psychometric instruments (e.g.,Prieler, 2007). However, there is little empirical evidence regarding the superiority of IRT in individual change assessment. In this study, we compared the CTT and IRT methods with respect to their sensitivity and specificity in detecting clinical and/or statistical significance of change in individual patients. In order to study the possible differences between CTT and IRT, we simulated polytomous items representative of clinical scales. Such items typically have high discrimination parameters and difficulty parameters which are concentrated at the upper part of the latent trait scale. Moreover, we also varied factors such the amount of true change and test length to test whether they interact with the methodology used for scoring the test (CTT vs. IRT) in correctly detecting change that is clinically and/or statistically significant. Preliminary results revealed differences between CTT and IRT, however, these differences were a function of the combination of different variables used in the study (i.e., scoring method, amount of true change and test length) rather than the scoring method alone.

Measuring Personality Using Conditional Reasoning

Paper Presentation

Lawrence R. James, *Georgia Institute of Technology*

Conditional reasoning is a measurement system designed to determine if people have particular types of biases that unknowingly affect their reasoning. The biases serve to rationalize societal sanctions against expressing motives such as a desire to harm others (aggression). People solve reasoning problems that have a bias built into one of the solutions. This solution appears logical to people who possess the bias. The solution is not, however, logically attractive to people who do not possess the bias. These people are offered a solution that appears more logical to them, such as reasoning based on prosocial principles for nonaggressive respondents. In other cases, the alternative solution is based on reasoning grounded in a contrasting bias. For example, in assessing dominance versus submissiveness, one solution is based on a bias that justifies a desire to dominate others, and another solution is based on a bias that protects submissiveness. This process is referred to as "conditional reasoning" because what is judged to be most reasonable is dependent on the motives and biases of the reasoner. This paper overviews the conditional reasoning process and summarizes some of its primary measurement characteristics, as described in a recent book by James and LeBreton (2012).

Performance Evaluation of a Two-tier Model under Variant Latent Trait Distributions

Paper Presentation

Hyesuk Jang, *Michigan State University*
Ji Seung Yang, *University of Maryland*

A two-tier model allows modeling correlated general factors while residual dependency among items is captured by orthogonal specific factors. One of the situations where a two-tier model can be adequately utilized is a longitudinal research design setting (Cai, 2010), in which a latent trait of interest is measured repeatedly. A two-tier model allows specifying dependency among the repeated items (or subset of items) as well as correlated latent traits across two time points. However, the distributional assumption of latent traits is not always met in some research contexts (e.g., when a treatment targets change in distribution of latent trait). The purpose of this study is to evaluate parameter estimation of a two-tier model through a simulation study when the distributions of the latent traits are non-normal but variant in term of skewedness and modes. Not only shapes of latent distributions but also magnitude of correlation between general factors, sample size, and the number of items will be varied in generating data. Using a normal distribution and an empirical histogram approaches to integrate out latent traits, item parameters are estimated and evaluated. Latent trait level estimates for general factors and their standard error estimates are also evaluated.

Detection of Differential Item Functioning using Recursive Partitioning

Paper Presentation

Heather Jeffers, *Ball State University*
Holmes Finch, *Ball State University*
Brian French, *Washington State University*

The detection of measurement invariance with differential item functioning (DIF) methodologies is used to provide evidence of score validity, with absence of DIF supporting construct validity. Several DIF methods, which use a single categorical conditioning variable, are effective for invariance assessment. A promising technique for DIF detection with multiple conditioning variables is a recursive partitioning tree approach based on the Rasch model. This Raschtree method can be used to detect DIF in multiple items simultaneously for both continuous (e.g. income) and categorical (e.g. gender) conditioning variables, simultaneously. Raschtree is grounded in model-based recursive partitioning, and is available in the R software library psychotree. In this approach, an initial Rasch model is fit for the entire sample, and is then tested to determine if item difficulty values differ for different levels of the partitioning variables (e.g. gender, income). If so, the sample is split by partitioning variables, and separate Rasch models are estimated for each partition. Partitioning is repeated until no further model differences are identified. In this simulation study, we assess the accuracy of Raschtree for DIF detection for different numbers of item, sample sizes, group size ratios, DIF magnitudes, and DIF contamination levels. Results support the effectiveness of Raschtree.

IRT Equating of Certification Test with Seasonality

Paper Presentation

Yanming Jiang, *Educational Testing Service*
Yuming Liu, *Educational Testing Service*

A certification test administered multiple times annually may display strong presence of seasonality. This seasonality is usually marked by the large variations in the percentage of first time test takers (FTTTs) across administrations such as in a high school exit examination. We seek to determine the optimal equating method for a certification test that exhibits strong seasonality. Based on a simulated target population with 70% FTTTs, subsamples with various percentages of FTTTs were analyzed. Three post-equating methods (unweighted, two-group, and weighted) and pre-equating method were evaluated. To determine the accuracy of the equating methods, equating estimates from the subsamples such as b-values, equated scores, and passing performance, were compared to those of the target population. When the percentage of FTTTs ranged from 40% to 60%, the weighted method had the most accuracy; when the percentage was 10%, the pre-equating method performed slightly better than the weighted method. When there were no FTTTs, the pre-equating method was better than the unweighted method based on repeaters. The unweighted method had the least optimal performance of the four methods considered. The optimal equating method may depend on the characteristics of an equating sample and could vary based on seasonality.

A Rasch Model for Carry-over Effect in Longitudinal Data

Paper Presentation

Kuan-Yu Jin, *The Hong Kong Institute of Education*
Wen-Chung Wang, *The Hong Kong Institute of Education*
Magdalena Mo Ching Mok, *The Hong Kong Institute of Education*

In many longitudinal studies, tests are administered repeatedly to measure growth across time. Such tests often consist of a set of common items to link all items on the same scale so that growth can be quantified. These common items are responded by the same persons more than once, which may result in carry-over effect. If so, the usual assumption of local independence is violated. If the carry-over effect exists but is ignored by fitting standard item response theory models, the parameter estimates will be biased and the conclusions will be misleading. To resolve this problem, we develop a new Rasch model that specifically account for the carry-over effect in common items in longitudinal data. Results of a series of simulations demonstrated that the parameters of the new model were recovered fairly well by using WinBUGS. An empirical example of growth in mathematical proficiency was provided to illustrate the implications and applications of the new Rasch model.

Exploratory Factor Analysis via Penalized Maximum Likelihood

Paper Presentation

Shaobo Jin, *Uppsala University*
Irini Moustaki,*London School of Economics and Political Science*
Fan Yang-Wallentin, *Uppsala University*

Maximum likelihood estimation is commonly used to conduct an exploratory factor analysis. Rotation methods are applied to the maximum likelihood factor loading matrix to achieve a sparse loading matrix for better interpretation. However, traditional rotation methods often produce a loading matrix with a few high loadings and a few small loadings. In this paper, we study the maximum likelihood estimation with an elastic net penalty term in the context of an orthogonal structure. The elastic net estimator shrinks the maximum likelihood estimator of factor loadings. An EM type algorithm is implemented to compute the elastic net estimator. A simulation study is conducted to investigate the properties of the elastic net estimator. The results show that the elastic net estimator is able to produce a sparse loading matrix with exact zero elements. The elastic net estimator is also compared with the other sparse estimation methods.

Modeling Response Styles Using Vignettes and Person-specific IRT

Paper Presentation

Katherine Grace Jonas, *University of Iowa*

An individual's response to a polytomous item is determined not only by item characteristics and the person's standing on the trait of interest, but also by the individual's use of the response options, and consequently, their response style. One recently proposed method for modeling response style with personality and attitudinal data involves the use of vignettes, in which respondents rate brief descriptions of targets at various levels of the trait of interest. Population-representative data from a polytomous measure of personality pathology (the Personality Inventory for DSM-5; PID-5) were modeled via a variant of generalized partial credit IRT with both item-specific and person-specific parameters. Model parameters were estimated using self and vignette ratings. The utility of estimating person-specific parameters was evaluated in terms the relationship among these parameters, updated estimates of personality traits, and a measure of adaptive functioning.

Development of Practical Rules for Identifying Equivalent Measurement Models in Structural Equation Modeling

Poster Presentation

Jung, Hyun Joo, *Sungkyunkwan University*
Lee, SoonMook, *Sungkyunkwan University*

In structural equation modeling, equivalent models can be observed for a given model, resulting in identical implied covariance matrices and thus identical overall fit indices. Equivalent models can be distinguished not by statistics of fit indices but by substantive interpretations. While previous research has been mainly concerned with equivalence in structural models, the present study will focus more on equivalence in measurement models. We will introduce several practical rules for deriving equivalent measurement models based on the concept of correlation (McDonald, 1985) and expanded JID approach (Lee, 1992, 2007; Lee & Kim, 2007) drawing on the property of a just-identified block in a structural equation model. In order to validate the rules, parameter mapping (Raykov & Penev, 1999) and computer simulation methods were used. In addition, applications of the practical rules will be demonstrated on empirical examples.

Multilevel Dynamic Generalized Structured Component Analysis for Brain Connectivity Analysis in Functional Neuroimaging Data

Paper Presentation

Kwanghee Jung, *University of Texas Health Science Center at Houston*

We extend Dynamic GSCA (Generalized Structured Component Analysis) to enhance its data-analytic capability in brain connectivity analysis of multi-subject functional neuroimaging data. Functional neuroimaging data are typically hierarchically structured, where time points are nested within subjects who are in turn nested within an experimental group. The proposed approach, named Multilevel Dynamic GSCA, explicitly accommodates the nested structure in functional neuroimaging data. Explicitly taking the nested structure into account, the proposed method provides fixed loadings between observed and latent variables in addition to fixed path coefficients between latent variables. Moreover, the proposed method allows investigation of subject-wise variability of the loadings and path coefficients by looking at the variance estimates of the corresponding random effects. We demonstrate the effectiveness of the proposed approach by applying the method to the two-level Sternberg working memory task data, where time series data-level measurements are nested within different subjects.

Considering the Bifactor Model

Paper Presentation

Walter Kaczetow, *CUNY Graduate Center*
Jay Verkuilen, *CUNY Graduate Center*
Ameer Mourad, *CUNY Graduate Center*

The bifactor model, proposed in the 1930s, has had a renaissance in the last ten years. It provides a parsimonious model when a general factor is of primary interest but local dependence is expected, e.g., due to symptom clusters or testlets. We consider some challenges posed by it. In particular, it seems to be prone to empirical underidentification compared to traditional CFA specifications. Upon reflection, this is perhaps unsurprising because the bifactors should be weak compared to the general factor for the model to be appropriate. Our approach involves considering the factor model as a restricted multivariate regression. If the bifactor model is sensible and the general factor is strong, we would anticipate collinearity problems. We propose to examine this through a simulation study. We also consider some diagnostics that may shed light on likely problems. In particular, we make use of the eigen-decomposition of the factor covariance matrix, the inverse of the factor covariance matrix, and the singular value decomposition of the structure matrix (pattern times factor covariance). Reasonable proxy values for these matrices can often be assessed before model fitting, which is likely to be particularly useful for nonlinear models such as binary or ordinal IRT.

Robust Bayesian IRT Modeling with Automatic Outlier Detection

Poster Presentation

Nicole Kaminski-Ozturk, *University of Illinois at Chicago*
George Karabatsos, *University of Illinois at Chicago*

Item response outliers may bias estimates of examinee ability and item parameters in IRT models. While fit statistics are used to identify outliers in IRT models, in practice they are used to justify the removal of items and/or persons from the analysis. Such practice is not uncontroversial. We propose a Bayesian approach to IRT modeling, where in addition to examinee and item parameters of the given IRT model, item response outlier indicator (0,1) parameters are added to the model. Specifically, when the item response is an outlier, its outlier parameter is 1, and then the model assigns a flat item characteristic curve (ICC) to that response; otherwise, when the item response is not an outlier, its outlier parameter is zero, and then the model assigns the response an ICC that is monotone increasing with examinee ability (e.g., a logistic ICC). The outlier parameters lead to estimates of the examinee and item parameters that are robust to the presence of outliers, and estimates provide model-based person and item fit statistics. Therefore, under this modeling approach, it is not necessary to remove items or persons from the analysis. Our IRT approach can lead to fast estimation of parameters via MCMC methods. We illustrate our approach through the analysis of real item response data, involving an IRT model with robust Student ICCs, being a function of examinee ability, item difficulty, and the outlier indicator parameters.

Online Calibration in Computerized Adaptive Tests at the Presence of Collateral Information in Response Times

Paper Presentation

Hyeon-Ah Kang, *University of Illinois at Urbana-Champaign*
Yi Zheng, *University of Illinois at Urbana-Champaign*
Hua-Hua Chang, *University of Illinois at Urbana-Champaign*

To more efficiently replenish an item bank for a computerized adaptive test (CAT), pretest item parameters can be calibrated during the operational CAT process (termed “online calibration”). The computer-based test delivery also enables the access to extra information on response times (RTs) specific for each test taker and individual items. The current literature has suggested that the information on RTs can improve the estimation of test takers’ latent abilities and adaptive item selection in CAT. However, little has been studied regarding the capitalization on RTs for online calibration. This study investigates the possibility of using collateral information in RTs for calibrating item parameters during the adaptive tests. A simulation study will be carried out to study new online calibration strategies based on RTs. The item responses are modeled under the three-parameter logistic model and the response times under a lognormal model. The item parameters of the pretest items are calibrated during the CAT within a hierarchical framework using an MCMC estimation approach. Findings of this study will provide implications for using collateral information in online calibration in which biases for item parameter estimates may be reduced.

A Comparison of MACS and RFA Models in Detecting Nonuniform Differential Item Functioning

Paper Presentation

Yujin Kang, *The University of Iowa*
Soonmook Lee, *Sungkyunkwan University*

This paper investigates the detection of nonuniform differential item functioning (DIF) by using mean and covariance structure (MACS) and restricted factor analysis (RFA) models, which are structural equation modeling (SEM) based methods, and compares their performance. The two models share the same assumptions related to general SEM, but they also have different characteristics. The RFA model implicitly assumes that the parameters in covariance matrix are identical across groups whereas the MACS model does not. Based on this difference, two hypotheses are set up. Hypothesis 1 is about the effect of the inequality of uniqueness between groups, and hypothesis 2 is about the effect of the inequality of sample sizes between groups on the performance of MACS and RFA models. The results partially support hypothesis 1 but do not support hypothesis 2.

Context Questionnaire Rotation and Imputation with Reference to Plausible Values in Large Scale Assessments

Paper Presentation

David Kaplan, *University of Wisconsin - Madison*
Dan Su, *University of Wisconsin - Madison*

In the context of international large scale assessments, it is desirable to obtain as much contextual information from respondents as possible. Such contextual information serves to place the assessment results in a larger policy context. However, a controversy exists regarding the desirability of rotating context questionnaires in large scale assessments because, it is argued, the conditioning models used to create the plausible values of the cognitive assessment require complete contextual questionnaire data. There is no consensus in the extant literature whether rotation is harmful to the creation of plausible values. Recently, however, the administration of PISA 2012 implemented a relatively simple rotation of the context questionnaire, and the public use data file does not contain imputations provided by the PISA consortium. Therefore, it is up to the user to implement a massive imputation. For this paper, we demonstrate, using PISA 2012, how the choice of imputation method can be validated, and more importantly, we show how the choice of imputation method can affect the estimation of models wherein the plausible values of the major cognitive domain are used as outcomes.

Pairwise Maximum Likelihood for Structural Equation Models: Estimation and Testing

State of the Art Talk

Myrsini Katsikatsou, *London School of Economics and Political Science*

Latent variable modelling is a well established framework within Social Sciences where the variables of interest are constructs, such as attitudes, beliefs, skills, etc. When the latent variables are taken to be continuous there are two approaches to latent variable modelling, the Underlying Variable approach used in Structural Equation Modelling (SEM) and the response function approach used in Item Response Theory. In both approaches, when the data are ordinal or of mixed type (ordinal and continuous), full information maximum likelihood (FIML) estimation is not feasible for large models. Within the SEM approach, limited information stepwise estimation methods have been developed and implemented in commercial software.

In this talk, we propose an alternative estimation method of likelihood type, the pairwise maximum likelihood (PML). The advantage of the method is that the whole machinery of FIML for inference can be extended to the case of PML. We develop a) a pairwise likelihood ratio test (PLRT) for testing nested structural equation models and overall goodness-of-fit of a model, and b) model selection criteria of AIC and BIC type under the pairwise likelihood framework. The latter are not possible to be estimated under the current three-stage limited information estimation methods. Simulation results will be presented to show the performance of PML estimation and the related test statistics under different simulation scenarios. Moreover, a comparison with existing methods of estimation and testing will be presented.

Stochastic Curtailment under Loglinear IRT

Paper Presentation

Henk Kelderman, *Vrije Universiteit*
Neils Smits, *Vrije Universiteit*

Stochastic Curtailment is increasingly used to make diagnostic decisions on the basis of total test scores without administering the full test. For example if one finds a very high probability of an at-risk decision based on the total score given subject's sum score on the first 10 response on a 20 item depression test, there is little need to administer the remaining 10 items. In principle the estimation of these conditional probabilities can be improved if a-priori information is available about the simultaneous distribution of the total and item scores. In this paper we model this distribution with quasi-loglinear models for the structurally incomplete contingency table of item responses and the total score. No contingency table is set up, but parameters are directly estimated from their sufficient statistics. Both expected sufficient statistics and conditional probabilities for various numbers of items administered are computed using efficient recursive formulae. Using cross validation we study the quality of diagnostic decisions under various models. We compare simple empirical conditional probabilities and estimates of these probabilities under the quasi-independence model, i.e. the loglinear Rasch model, and its extensions with a Markov structure of consecutive item responses of various orders. The methods are illustrated analyzing simulated and empirical data.

Neural Networks and Random Forests for Propensity Score Estimation

Paper Presentation

Bryan Keller, *Teachers College, Columbia University*

Propensity scores (PSs) are most often estimated by logistic regression. If the relationship between the PS and the covariates is complex, the logistic model must include interactions and other higher-order terms to capture nonlinearities in the selection surface. Data mining techniques that algorithmically handle nonlinear relationships and interactions have been noted as promising for PS estimation because they deal with such nonlinearities in their basic implementation. In this talk, we examine the performance of two such methods (neural networks and random forests; NNs and RFs) for PS estimation via simulation and compare their performance to a baseline logistic model. When the data-generating PS model was simple (i.e., only contains main effects terms) NNs and RFs performed comparably. However, when the data-generating PS model also contained quadratic and second-order interactions, NNs were better able to provide covariate balance on those terms. We also observe that the PS estimation approach which resulted in the best balance on main effects covariate terms was not always the one which produced the least biased treatment effect estimates. These results highlight the importance of checking balance on more than just first-order covariate terms.

Multiple-groups Measurement Invariance

Paper Presentation

Justin L Kern, *University of Illinois at Urbana-Champaign*
Brent A. McBride, *University of Illinois at Urbana-Champaign*
Daniel J. Laxman, *University of Illinois at Urbana-Champaign*
W. Justin Dyer, *Brigham Young University*

Measurement invariance (MI) is a property of measurement that is often assumed, but rarely tested. Standard techniques for investigating MI have been developed for the case of two groups. However, very little work has been done on the case of more than two groups, even though the need for such techniques is apparent in many fields of research. This paper introduces and illustrates a model building technique to investigating MI for more than two groups. This technique is an extension of the already-existing hierarchy for testing MI introduced by Meredith (1993). An example using five different groups of families of children with and without disabilities from the Early Childhood Longitudinal Study-Birth Cohort (ECLS-B) dataset will be given.

Using Data Mining Techniques to Extract Substantively Meaningful Features from Process Data

Paper Presentation

Deirdre Kerr, *ETS*

I will be discussing my work using data mining techniques to extract substantively meaningful features from process data from an educational video game called Save Patch. Save Patch was designed to teach students about the identification of fractions using a 'gamified' number line representation. The game is broken into stages of increasing mathematical complexity, each consisting of three to five levels addressing the same content area. To use the game as an assessment of student understanding, we first had to develop methods of extracting substantive information from the game's process data. The first step of this process was to try to identify the strategy each student was using in each attempt to solve each level in the game. I will discuss how fuzzy feature cluster analysis was used to group individual actions in the game into interpretable strategies, allowing for the identification of specific misconceptions and gaming behavior, as well as a variety of valid solution strategies. I will then discuss how I identified changes in strategy use that might indicate learning, confusion, or mastery by integrating expert knowledge and sequence mining.

Exploring Simpler Alternatives to the 3-parameter Logistic Model for Operational Use

Poster Presentation

EunHee Keum, *The Ohio State University*
Michael C. Edwards, *The Ohio State University*

The 3-parameter logistic model (3PLM) is a commonly used item response model in high-stake educational testing, which attempts to 'correct' for the possibility of guessing the correct answer on a multiple choice question. However, it can be very challenging to use in practice due to technical difficulties in obtaining a stable solution. Additionally, the interpretation of the parameter and the impact it has on individual examinees are not very clear. These issues suggest it may be worth more seriously considering simpler competitors such as the 2-parameter logistic model (2PLM) and the 3PLM with a fixed lower asymptote. The purpose of this study was to understand the implications of choosing different models to deal with multiple choice data. By using the operational data from large operational tests and targeted simulations, we evaluated the performance of some different, but simpler, competitors to the 3PLM. We evaluated the impact of the choice of these different models on calibration, scale construction, scoring, and stability of all model-based activities. Furthermore, we attempted to develop a knowledge base of issues surrounding operational use of the 3PLM in the high-stake testing and possibly provide guidelines.

Behavioral Analytics in Intelligent Training Systems

Paper Presentation

Saad Khan, *Educational Testing Service*
Alina von Davier, *Educational Testing Service*

As the boundary blurs between what is real and what is virtual in today's educational/training environments, there is a growing need for new assessment tools that capture behavioral aspects key to evaluating skills like problem solving, communication and collaboration. This talk presents an approach for developing interactive training systems that capture and analyze trainee/s behavior at multiple levels of abstraction for automated performance assessment. A key challenge considered here is to capture and understand human (trainee) behavior at fidelity sufficient to estimate trainee's cognitive and affective state as manifests through multiple mediums including speech, body pose, gestures, gaze etc. However, analyzing each of these modalities in isolation may result in incongruities. In addition, the affective states of a person show significant variations in time. To address these issues we model the temporal dynamics and integration of multiple data modalities using conditional random fields (CRFs) on observable data. We demonstrate our approach in the context of simulated intelligent tutoring systems where a detailed understanding of human/trainee behavior is used to customize and drive the flow of training scenarios. We also demonstrate how this approach may find application in assessment, particularly in virtual learning environments.

Measuring and Correcting for Response Styles in Large Scale Assessments using HYBRID Models and a New MIRT Approach

Paper Presentation

Lale Khorramdel, *Educational Testing Service*
Matthias Von Davier, *Educational Testing Service*

The current study introduces a multidimensional item response theory (MIRT) approach to measure and correct for response styles (RS) in rating scales. Based on an approach presented by Böckenholt (2012), responses to a 5-point rating scale are decomposed into multiple response sub-processes and modeled through MIRT models to examine the extreme and midpoint RS. Using a HYBRID model (Yamamoto, 1989) extended to polytomous response formats (von Davier, 1996), the investigated sample is divided in three latent classes: an IRT class which is assumed to provide information about the skill level, a zero inflated count class showing a persistent pattern of skill non-use, and a response style class which is assumed to give construct irrelevant responses. The MIRT approach is carried out along side with the HYBRID model and the results are compared. The data come from a background questionnaire of an international large scale assessment measuring different scales, which are believed to correlate with the cognitive part of the assessment. It is shown that RS can be measured as unidimensional and differentiated from trait-related responses. Results are discussed with regard to group differences, and with regard to the influence of RS on the correlation between background variables and cognitive test scores.

The Relationship Between Cooperative Learning in Mathematics and Students? Self-confidence in Learning Mathematics, How They Value It, and Their Positive Affect Toward It

Paper Presentation

Daeseok Kim, *Kongju National University*
Hyungjung Cho, *Soonchunhyang University*

This study investigated the relationship between cooperative learning in mathematics and students' self-confidence in learning math (SCM), attitude toward it (PATM), and the value they place on it (SVM). The results show that students who work together in small groups in all or almost all mathematics lessons have more positive attitudes (SCM, PATM, and SVM) toward the subject than those who never work together. The differences between the two samples were all statistically significant regardless of gender, mathematics score, and parent education level. Therefore, the students' attitude toward mathematics is presumed to improve when they work together in small groups in during the lessons, regardless of their background. Since the biggest difference between the two samples was in PATM, small group cooperative learning in mathematics is presumed to have the strongest impact on PATM. The difference between students whose parents did not attain the upper secondary school level were bigger than those whose parent education level was a college graduation or higher. Therefore, small group cooperative learning in mathematics is presumed to make a heavier impact on students whose parent education level is low. The results were consistent in both the United States and South Korea.

The Effect of Ignoring Event Dependency in Event-history Analysis: An Illustration with the Early Steps Observational Data

Paper Presentation

Hanjoe Kim, *Arizona State University*

There can be two sources of dependency while analyzing parent-child dyads event-history (survival) data. One comes from the multilevel structure of the data. Hazard rates of recurring events within a parent-child dyad can be more similar to each other than repeated measures from other parent-child dyads. The other source of dependency can come from a bundle of competing events. An increase of a certain event (hazard) rate implies a decrease of competing event rate(s). Ignoring the event dependency can cause biased inference (Dagne & Snyder, 2009). In a previous study (Kim, Dishion & Tein, 2014), a Cox proportional hazards model with shared frailty was used to analyze the Early Steps (Dishion & Stormshak, 2007) observational data. The termination and return time of two out of five dyadic states were analyzed. However, the event dependency was not considered in the previous study. As an extension of Dagne and Snyder (2009), a method of incorporating the competing risks in a more general SEM framework is proposed and applied to the Early Steps data. Results from the model with and without the competing risks are compared. Future directions and applications are discussed.

An Exploration of Factors Predicting Achievement Profiles and Score Report using Cognitive Diagnostic Model

Paper Presentation

HeeKyoung Kim, *Korea Institute for Curriculum and Evaluation*
Jung-A Han, *Korea Institute for Curriculum and Evaluation & Yonsei University*

Variables affecting CDM-based achievement profiles were explored using an SEM approach. Innovative ways for score reporting of the CDM analysis results are explored to provide individual students.

Multilevel Propensity Score Methods for Estimating Causal Effects: Challenges & Strategies

Paper Presentation

Jee-Seon Kim, *University of Wisconsin-Madison*
Peter M. Steiner, *University of Wisconsin-Madison*

There has been a strong and increasing interest in propensity score analysis as a tool for making causal inferences in quasi-experimental and observational research in the social and behavioral sciences. Although the standard use of propensity score methods is rather well established, many aspects of the methods are less well understood, particularly in the context of clustered or nested data structures. This talk focuses on fundamental issues in evaluating treatment effects from multilevel observational studies and suggests strategies for making proper inferences. Specifically, we demonstrate how to incorporate different selection mechanisms across different clusters and how to deal with a lack or absence of overlap between treated and non-treated groups within clusters. When there is a lack of overlap within clusters, one can exploit the multilevel structure of the data and borrow cases from other clusters, but several issues need be considered in terms of which clusters and which cases should be used for these purposes. With illustrative examples, this talk explains challenges, strategies, and diagnostics in multilevel propensity score methods, and also demonstrates consequences when these issues are ignored or improperly accounted for in data analysis.

Diagnosing Examinees' Attribute-Mastery using the Bayesian Inference for Binomial Proportion: A New Method for Cognitive Diagnostic Assessment

Paper Presentation

John H. Kim, *Vanguard University of Southern California*
Susan E Embretson, *Georgia Institute of Technology*
Mia O Kim, *Vanguard University of Southern California*

Although cognitive diagnosis models (CDM) have successfully provided useful diagnostic information about the examinee, most CDMs are complex due to a large number of parameters in proportion to the number of skills (attributes) to be measured in an item. The large number of parameters causes heavy computational demands for the estimation. Also, a variety of specific software applications is needed depending on the chosen models. The purpose of this study was to propose a simple and effective method for cognitive diagnostic assessment (CDA) without heavy computational demand using a user-friendly software application. Bayesian inference for binomial proportion (BIBP) was applied to CDA. The application of BIBP to CDA can be flexible depending on the test item-attribute design and examinees' attribute-mastery patterns. In this study, effective ways of applying the BIBP method was explored using real data studies and simulation studies. Also, other preexisting diagnosis models (e.g., DINA, LCDM) were compared to the BIBP method in their diagnosis results. BIBP appeared an effective method for CDA with the advantage of easy and fast computation and a relatively high accuracy of parameter estimation.

Comparison of Item Parameter Estimates Under Various Item Calibration Methods for a Two-Parameter Logistic Model

Paper Presentation

Kyung Yong Kim, *The University of Iowa*
Won Chan Lee, *The University of Iowa*

The purpose of this study is to investigate the effect of various item calibration methods on item parameter estimates for a two-parameter logistic model using a simulation study. Widely used item response theory (IRT) programs incorporate the marginal maximum likelihood estimation approach of Bock and Aitkin (1981) or the marginalized Bayesian estimation procedure of Mislevy (1986) to estimate item parameters. However, the numerical integration method and the latent variable distribution assumption differ from program to program. This study compares four item calibration methods: (a) using the midpoint rule for numerical integration with the standard normal latent distribution; (b) using the midpoint rule for numerical integration with a normal latent distribution; (c) using the midpoint rule for numerical integration with a posterior latent distribution; and (d) using the Gauss-Hermite quadrature rule for numerical integration with the standard

normal latent distribution. These four methods will be compared under various simulation conditions including different sample sizes, different number of items, and the use of prior distributions for item parameters. Results of this study may provide useful information about the relative performance of the four item calibration methods in estimating item parameters for a two-parameter logistic model.

Model Selection in Random Item Mixture IRT Models

Paper Presentation

Meereem Kim, *University of Georgia*
Hye-Jeong Choi, *University of Georgia*
Youn-Jeong Choi, *University of Georgia*
Allan Cohen, *University of Georgia*

Algorithms for estimation of IRT model parameters typically treat items as fixed and persons as random. This approach also is common in mixture IRT models. A more theoretically appealing approach is one in which items are considered as random rather than fixed, as items are typically considered to be drawn from a population. Further, in a random item model (RIM) uncertainty in item parameters can be included (De Boeck, 2008). Some model selection research has been reported for fixed item mixture IRT models (Li Cohen, Kim & Cho, 2009), but research on model selection for mixture RIM models is needed. In this study, a simulation is conducted comparing model selection indices for use with fixed and RIM mixture IRT models. Information indices including Akaike's information coefficient (AIC) and Bayesian information coefficient (BIC) are considered along with Bayes factors, and sample adjusted BIC. Test length (15 and 30 items) and sample size (800 and 1,600 examinees) conditions are manipulated. Results from a preliminary study indicate that AIC and BIC both tend to select the correct mixture RIMs, but AIC tends to select the more complex fixed item mixture models. Determining the penalty terms for different information indices will be examined.

Gauss-Hermite Quadrature in Marginal Maximum Likelihood Estimation of Item Parameters

Paper Presentation

Seock-Ho Kim, *The University of Georgia*
Yu Bao, *The University of Georgia*
Erin Horan, *The University of Georgia*
Meereem Kim, *The University of Georgia*
Allan S. Cohen, *The University of Georgia*

Although many theoretical papers on the estimation method of marginal maximum likelihood of item parameters for various models under item response theory mentioned Gauss-Hermite quadrature formulae (e.g., Bock & Lieberman, 1970; Bock & Aitkin, 1981), almost all computer programs that implemented marginal maximum likelihood estimation employed other numerical integration methods (e.g., Newton-Cotes formulae). There are many tables that contain quadrature points and quadrature weights for the Gauss-Hermite quadrature formulae (e.g., Abramowitz & Stegun, 1972; Stroud & Secrest, 1966); but these tabled values cannot be directly used when quadrature points and quadrature weights are specified by the user of computer programs because the standard normal distribution is frequently employed in the marginalization of the likelihood. There are two purposes of this paper. The first one is to present extensive tables of Gauss-Hermite quadrature for the normal distribution mainly based on the computer program from Stroud and Secrest (1966). The other is to present an example that demonstrates the effect of using various numbers of quadrature points and quadrature weights as well as different quadrature formulas on item parameter estimates. Item parameter estimates obtained from more than 20 quadrature points and quadrature weights with either Gauss-Hermite quadrature or Newton-Cote quadrature were virtually identical.

The Impact of Measurement Scale and Nonnormality of Indicator Variables in Latent Growth Models

Poster Presentation

Su-Young Kim, *Ewha Womans University*
Youngsuk Suh, *Rutgers University*

It has been a quite common practice in psychology or the social sciences that we treat ordinal variables as continuous in various statistical models, such as factor analysis, if the variables have many categories (e.g., five). Some empirical evidence has been provided that treating ordinal variables with many categories as continuous does not have much practical impact on statistical results (see e.g., Babakus, Ferguson, & Jöreskog, 1987). For a latent growth model (LGM), substantive researchers typically do the same practice when there are not multiple indicators for a latent construct or when they want to directly model a single manifest variable. The impact of measurement scale and nonnormality of indicator variables in LGM is examined in the present study. In particular, the impact of six kinds of indicator conditions on the estimation quality are investigated through a series of Monte Carlo simulations. Because the estimation quality of growth models is closely related with the sample size and the number of indicator variables (Kim, 2012), these two are also crossed with the indicator variable conditions. The results of the present study will provide particularly useful information and recommendations to substantive researchers who plan to use LGMs with ordinal variables.

Measurement of Emotional and Behavioral Disorders in Adolescents by Using Latent Classification Models

Poster Presentation

Sung Eun Kim, *Ewha Womans University*
Seul Ki Koo, *Ewha Womans University*

The present study aims to make a diagnosis of emotional and behavioral disorders in adolescents based on latent classification models. The present study performed 3 steps in the research process. In the first step, we develop the Q-matrix. The First draft of the Q-matrix is constructed through the experts' discussion (or specialists' conferences) and statistical Q-matrix validation is applied to revise the Q-matrix. The 5 experts discuss to decide the final Q-matrix. As the second step, analysis is performed using the G-DINA (generalized deterministic input; noisy "and" gate) model. In the last step, we interpret the results and provide an example in which we propose a new form of reporting the results to present a diagnosis of psychological problems. In this study, the response data of The Korean Children and Youth Panel Survey (KYCPS) were used. KCYPS selected 21 questions constructed of 3 factors (attention deficit and hyperactivity, aggressive behavior, somatoform) and were collected responses targeting 2,351 people the eighth grade. In the results of the study, we can investigate the behavioral disorders that have the highest retention rate of youth in Korea. Furthermore, we hope that CDMs will become more widely used in the psychological field in the future as well.

A Latent Class Item Response Theory Model

Paper Presentation

YoungKoung Kim, *The College Board*
Lawrence T. DeCarlo, *Teachers College, Columbia University*

Many large scale assessments report examinee ability in terms of classifications – for example, three levels (Basic, Proficient and Advanced) for the NAEP and five levels (from Extremely well to No recommendation) for the AP®. These classifications are often used to make decisions about enrollment in remedial courses, granting college credits etc. In practice, in order to arrive at such classifications, two steps are required: 1) examinee responses are scored using some form of continuous scoring, such as item response theory (IRT) ability scores, and 2) standard setting procedures are then used to determine cut scores for the continuous scores. The present study proposes a latent class approach in order to directly obtain examinee classifications. Instead of assuming continuous latent examinee ability, as in IRT, the present approach assumes ordinal latent classes of examinee ability. This can be done to directly obtain the (3 or 5 level) examinee classifications. The proposed model is an extension of a latent class (LC) model that has been used for essay grading (DeCarlo, 2005). Mixed-format large scale assessment data are used to illustrate the LC IRT model. Classification accuracy for the LC model is compared to that obtained for the IRT model (with cut-points).

Bias-Amplification in Observational Studies

Paper Presentation

Yongnam Kim, *University of Wisconsin-Madison*
Peter M. Steiner, *University of Wisconsin-Madison*

In observational studies, differential selection into treatment typically results in treatment and control groups that are not directly comparable. Nonetheless, causal average treatment effects are identified if the selection mechanism is ignorable. In order to establish ignorability, researchers running propensity score or regression analyses frequently try to control for as many covariates as possible in the hope to cover all relevant confounders or at least their correlates. In our talk, we demonstrate that controlling for a confounder actually has two opposed effects on the bias: a bias-reducing effect that removes overt bias and a bias-amplifying effect that amplifies any hidden bias (if ignorability is not given). Using analytical formulas and simulation studies, we show how near instrumental variables (these are confounders that are strongly related to the treatment but are only weakly related to the outcome) affect the imbalance in unobserved confounders and, thus, amplify the bias in the treatment effect. We also show how the impact of near instruments varies as a function of the confounders' heterogeneity, correlation structure, and measurement reliability. We then investigate under which conditions the bias-amplifying effect dominates the bias-reducing effect and extend the findings to multilevel settings.

A Finite Mixture of Nonlinear Random Coefficient Models for Continuous Repeated Measures Data

Paper Presentation

Nidhi Kohli, *University of Minnesota*
Jeffrey R. Harring, *University of Maryland*
Cengiz Zopluoglu, *University of Miami*

Nonlinear random coefficient models (NRCMs) for continuous longitudinal data are often used for examining individual behaviors that display nonlinear patterns of development (or growth) over time in measured variables. As an extension of this model, this study considers the finite mixture of NRCMs that combine features of NRCMs with the idea of finite mixture (or latent class) models. The efficacy of this model is that it allows the integration of intrinsically nonlinear functions where the data come from a mixture of two or more unobserved subpopulations, thus, allowing the simultaneous investigation of intra-individual (within-person) variability, inter-individual (between-person) variability and subpopulation heterogeneity. In this article the efficacy of this model to work under real data analytic conditions was examined by executing a Monte Carlo simulation study. The simulation study was carried out using an R routine specifically developed for the purpose of this study. The R routine used Maximum Likelihood (ML) with the Expectation-Maximization (EM) algorithm. Finally, the utility of the model is illustrated using a real data example.

Comparison of Common Item Scale Transformation Methods through Non-Equivalence between Test Difficulty as well as between Examinee Class

Poster Presentation

Seul Ki Koo, *Ewha Womans University*
Sung Eun Kim, *Ewha Womans University*
Inyoung Park, *Ewha Womans University*
Jiae Pyun, *Ewha Womans University*

Vertical scaling is carried out when understanding development of subjects' ability or comparing same grade scores through the years. On the vertical scale, a difference between grades can induce a distinction of ability distribution; test construction adjusted to target grades' ability can induce a distinction of average difficulty between tests. The scale transformation for comparing the same grades also cannot assure that the subjects' capability level is similar every year. The most correct ability parameter can be computed when each grade's capability level and test's difficulty is organized appropriately; but it can cause the difference of ability between the examinee classes and the difference between tests on the scale transformation at the same time; this would be a factor that hampers the scale transformation's accuracy and stability. Therefore this research employs separate calibration, which is the one of the scale transformation methods using anchor items in order to search for a more stable way of scale transformation; and compares its stability and accuracy (RMSE, BIAS, SEE) through a simulation study applied using Stocking-Lord and Fixed Item Parameter Calibration. The study can help understand each scale transformation method in a common school situation and provide valuable information on choosing appropriate methods.

A Study on Improving the Reliability and Validity of Scoring in the NEAT

Poster Presentation

Seulki Koo, *Korea Institute for Curriculum and Evaluation*
Yongsang Lee, *Korea Institute for Curriculum and Evaluation*

In the present study, we examined various issues related to scoring reliability and validity in the National English Ability Test (NEAT) in order to suggest effective ways to improve reliability and validity in scoring. For this purpose, this study explored various factors which may affect the scoring reliability and validity, and analyzed the rater training programs and scoring procedures of foreign testing agencies. Through in-depth interviews with the NEAT raters, this study also tried to examine the relationships among rater characteristics, rater training, scoring procedures, and the scoring reliability and validity. Based on the analysis results, the current study tried to discover ways to improve the rater training programs, scoring procedures, and scoring guides for the NEAT, and also suggested scoring strategies for performance assessment in schools.

Assessing Factorial Invariance of Two-Way Rating Designs using Three-Way Methods

Paper Presentation

Pieter M. Kroonenberg, *Leiden University*

Assessing the factorial invariance of two-way rating designs such as ratings of concepts on several scales by different groups requires three-way models such as the Parafac and Tucker models. By definition these models require double metric factorial invariance. Differences between these models are contained in their handling of the interactions between the concept and scale spaces. The interactions may consist of unrestricted linking (Tucker2 model), invariant component covariances but variable variances per group and component (Parafac model), zero covariances and variances different per group but not per component (Replicated Tucker3 model) and strict invariance (Principal component analysis on the average matrix). This hierarchy of invariant models and procedures to evaluate the models against each other, will be illustrated with an international data set from attachment theory.

Multiple Group Factor Analysis Model for Multitrait-multirater Data to Assess the Reliability and Validity Regarding the Number of Raters

Poster Presentation

Saori Kubo, *Waseda University*
Hideki Toyoda, *Waseda University*

The multitrait-multimethod (MTMM) matrix can be utilized as a means to examine the convergent and discriminant validity of psychological measurement. When MTMM approach is applied to human resource assessment, the traits typically represent performance dimensions, and the methods refer to different rating sources, thus within this context MTMM becomes MTMR (multitrait-multirater). In our study, the confirmatory factor analysis (CFA) model is applied to MTMM data to assess the reliability and validity. Because the traits and the methods are treated as trait factors and method factors respectively in a CFA model, the number of raters contained in the same rating source, if any, is ignored, and usually the mean of the rating value within the rating source is used as observed variable. The number of raters is supposed to affect the reliability and validity of measurement, but the conventional confirmatory factor analytic approach cannot assess the influence. Therefore, in order to accurately assess the effect of the number of raters on reliability and validity, this study shows a multiple group factor analysis model for MTMR data.

Model fit assessment in multilevel exploratory factor analysis

Paper Presentation

Megan Kuhfeld, *University of California, Los Angeles (UCLA)*
Li Cai, *University of California, Los Angeles (UCLA)*

We propose to examine the effect of model error on factor recovery and model fit indices in a two-level exploratory factor model. Given that researchers are increasingly implementing multilevel factor models to investigate the dimensionality of empirical hierarchical data, it is important to know whether widely-used tools for model fit assessment are sensitive to misfit in this context. Data will be generated from a multilevel factor model with three true factors at each level, and model misfit is introduced at one level of the factor model through ten additional weak minor factors following the Tucker, Koopman & Linn (1969) procedure. Various conditions will be examined to see if goodness-of-fit tests and fit indices (RMSEA, CFI, TLI) that are well-established with single-level data lead to accurate decisions about dimensionality when fitting multilevel EFAs. Additionally, this study will investigate the use of a recently-suggested Posterior Predictive Model Checking (PPMC)-analogous approach as an alternative to existing fit indices (Lee, Cai & Kuhfeld, under review). It is hypothesized that traditional methods of model fit evaluation will not be sufficiently informative with multilevel data, and that this new alternative approach will provide more useful and accurate tools for researchers examining the dimensionality of their data.

Bias in Estimates and Standard Errors of Mokken's Scalability Coefficients

Paper Presentation

Renske E. Kuijpers, *Tilburg University*
L. Andries van der Ark, *University of Amsterdam*
Marcel A. Croon, *Tilburg University*

In Mokken analysis, scalability coefficients are used both as criteria for item partitioning and as diagnostics for the strength of the scales. There are three types of scalability coefficients: (1) for pairs of items, (2) for items, and (3) for the entire scale. Recently, we used the marginal modeling approach to derive standard errors for those scalability coefficients. The estimates and standard errors of the scalability coefficients are derived assuming that the ordering of the item steps in the sample is identical to the ordering of the item steps in the population. If this assumption is violated, the estimates and standard errors may be biased. By means of a simulation study we investigated the bias of the estimates and standard errors of the scalability coefficients, and we assessed the coverage of the confidence intervals. In the simulation study, different factors - like item ordering, sample size, number of items, number of response categories - were varied.

Empirical and Exemplary Dataset Methods for Estimating Statistical Power in Longitudinal Designs

Poster Presentation

Kevin A. Kupzyk, *University of Nebraska Medical Center*

Several different methods of estimating statistical power are available. Formulaic power functions, such as those used in G*Power or Optimal Design, do not allow the user to account for realistic conditions such as attrition, unequal cluster sizes, or unequal group sizes. The simulation-based and exemplary dataset methods both allow complex situations to be accounted for, although little is known about the exemplary dataset method. A disadvantage of simulation is that it is computationally intensive, often requiring tens of thousands of replications in order to come to a stable estimate of power at each sample size attempted. O'Brien and Muller (1993) developed the exemplary dataset method based on the idea that the non-centrality parameter for a test would be the test statistic obtained if the group means and variances are equal to the population values. This analysis shows how this method is performed and how it compares to simulation. The two provided equal values, although the empirical method had to be performed with up to 50,000 replications in some cases to come to the same result as the exemplary method. Thus, the exemplary method may be helpful in determining power more quickly and precisely than by using simulations.

Nominal Response Model for Scoring Situational Judgment and Other Personality Tests

Paper Presentation

Patrick C. Kyllonen, *Educational Testing Service*
Jiyun Zu, *Educational Testing Service*
Hongwen Guo, *Educational Testing Service*

Situational judgment tests (SJTs) are commonly used in education and workforce contexts to measure respondents' personality, emotional intelligence, and noncognitive skills, such as leadership, teamwork, and communication skills. A common SJT item type comprises a situation description followed by a set of 4-7 responses to the situation with the examinee being asked to choose the best and worst of those options. Because SJTs are often designed to measure subtle qualities in judgment there is not necessarily a clear best and worst option, nor an ordering of options. In this paper we show that the nominal response model provides a good framework for scoring SJT items, particularly with ambiguous keys. We also show that option characteristic curves (Ramsay, 1991) provide useful diagnostic information to motivate NRM scoring. We compare NRM scores with other scores (number correct and generalized partial credit model, based on most frequent responses, and "consensus scores" which give partial credit in proportion to the sample choosing a particular option) and find that NRM scores have higher reliability and higher correlations with construct-related external variables. We discuss extensions to other kinds of noncognitive tasks with ambiguous keys.

Markov Decision Processes for Making Student Inferences from Complex Task Data

Paper Presentation

Michelle LaMar, *UC Berkeley*

Complex computerized tasks provide data not only on the final task outcomes, but also on the processes by which students arrived at their results. These data appear to provide valuable information about student thinking but are challenging to model with traditional psychometric approaches. This talk will explore the use of a Markov decision process (MDP) as a cognitive model of students working through a complex task. The MDP, when combined with an IRT approach, can enable inference about student ability from within-task process data. A simulation of a simple peg solitaire game will be used to illustrate the estimation and interpretation of model parameters. Parameter recovery will be examined under conditions of varying task complexity and varying student motivation. Application to a real educational game about cell biology will demonstrate both the feasibility and complexities of using this approach 'in the wild'. Additionally, issues of model identifiability and inference validity will be discussed.

An Empirical Investigation of Different Estimation Methods for the 3PL IRT Model

Paper Presentation

Sunil Lamsal, *Southern Illinois University Carbondale*
Yanyan Sheng, *Southern Illinois University Carbondale*

Different estimation procedures have been developed for the unidimensional three-parameter item response theory (IRT) model. These techniques include the Marginal Bayes estimation, the fully Bayesian estimation using the Markov chain Monte Carlo simulation techniques, and the Metropolis-Hastings Robbin Monro estimation. With each technique, a prior can be specified to reflect prior belief on each model parameter. Previous studies evaluating the fully Bayesian estimation procedure for this model suggests that the model can be viewed as a mixture model, and it suffers from a nonconvergence problem unless strong informative priors are specified for the item slope and intercept parameters. This study focused on comparing the three estimation methods for the three-parameter logistic (3PL) model in parameter estimation using Monte Carlo simulations. In particular, sample sizes, test lengths and prior specifications were manipulated to reflect various test situations. Results suggest that the three methods performed similarly with a slight advantage to the fully Bayesian estimation. Similar to findings with previous studies, a relatively more informative prior had to be specified for item parameters to ensure convergence with each of the three methods, and when sample size and/or test length was large, the model parameters were more accurately estimated.

Time-varying Learning and Content Analytics via Sparse Factor Analysis

Poster Presentation

Andrew Lan, *Rice University*
Christoph Studer, *Cornell University*
Richard Baraniuk, *Rice University*

We propose SPARFA-Trace, a new machine learning-based framework for time-varying learning and content analytics for educational applications. We develop a novel message passing-based, blind, approximate Kalman filter for sparse factor analysis (SPARFA) that jointly traces learner concept knowledge over time, analyzes learner concept knowledge state transitions (induced by interacting with learning resources, such as textbook sections, lecture videos, etc., or the forgetting effect), and estimates the content organization and difficulty of the questions in assessments. These quantities are estimated solely from binary-valued (correct/incorrect) graded learner response data and the specific actions each learner performs (e.g., answering a question or studying a learning resource) at each time instant. Experimental results on two online course datasets demonstrate that SPARFA-Trace is capable of tracing each learner's concept knowledge evolution over time, analyzing the quality and content organization of learning resources, and estimating the question--concept associations and the question difficulties. Moreover, we show that SPARFA-Trace achieves comparable or better performance in predicting unobserved learner responses than existing collaborative filtering and knowledge tracing methods.

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

Paper Presentation

Ashley Lawrence, *University of Oklahoma*
Taehun Lee, *University of Oklahoma*

It has been known in theory that two classes of missing data procedures, maximum likelihood (ML) and multiple imputation (MI) are equivalent. It has also been known empirically that ML and MI are not equivalent as practiced. However, there has been a surprising lack of empirical research into the relative performance of ML and MI in the context of structural equation modeling (SEM). A primary goal of this paper was to examine conditions in which ML and MI yield equivalent or divergent results. Specifically, we were interested in how ML and MI procedures perform when the model used in data analysis did not hold in the population. Monte Carlo studies were designed to examine the impact of various independent variables, including missing percentage, sample size, and degree of model misspecification on the outcome measures, including convergence rates of model estimation and bias and root-mean-squared-error for parameter and standard error estimates. Of particular interest was the discrepancy between parameter estimates estimated from the complete data and those estimated from MI and ML on incomplete data. Preliminary results indicate that the degree of model misspecification has important implications for the relative performance and equivalence of MI and ML.

Evaluation of Item Parameter Recovery Estimation by ACER ConQuest Software

Paper Presentation

Luc Le, *Australian Council for Educational Research*

This study used Monte Carlo simulations to evaluate the item parameter recovery from ACER ConQuest software (Adams, Wu, & Wilson, 2012; Wu, Adams, & Wilson, 1997) for the dichotomous Rasch model. Our primary focus was the comparison of its estimation methods, joint maximum likelihood (JML), marginal maximum likelihood (MML) with a normal distribution assumption and MML with a discrete distributions assumption when the populations were in fact non-normal. The simulation data sets were generated with two test lengths (10 and 50 items) and four alternative true population distributions for the abilities: normal, bimodal, uniform, and chi-square. As expected, results showed that MML-Normal was the best method when the assumption of ability distribution was matched, regardless the test length. However, the accuracy of MML-Normal decreased with the violation level of the assumption of normal distribution of the latent ability. The MML-Discrete estimation could overcome well the weakness of the MML-Normal when the normality of the ability distribution was violated. The estimates of the corresponding standard errors produced by ACER ConQuest were also being examined and discussed.

Predictive Validity of the Preschool Children's Learning and Development Indicators

Paper Presentation

Mei Ling Le, *Beijing Normal University*
Tao Xin, *Beijing Normal University*

This study aims to evaluate the predictive validity of the indicators of the preschool children's learning and development, which means to assess the effectiveness and fairness of indicators developmental prediction. The tools include the six areas which are physical and motor development, social-emotional development, cognitive development, language communication development, art appreciation and performance development, and approach to learning. A national representative sample of 3577 3-6 years-old children (of roughly equal percentages of girls and boys) was obtained who respectively, from the urban and rural of the East, Central, Western region. The fairness test was to detect differential item functioning (DIF) for testlets by gender, urban and rural, region using four procedures (p-MH, p-LR, p-SIBTEST and IRT-LR) and then to compare the results across the procedures in order to choose the non-DIF items. Consistent results were produced between the p-MH, p-LR, and p-SIBTEST but not the IRT-LR. Most items can reflect the children's developmental trends, indicating good predictive validity, and with the better performance, children grow better.

Detecting Significant Intraindividual Change with Conventional and Adaptive Tests

Paper Presentation

Ji Eun Lee, *University of Minnesota*
David J. Weiss, *University of Minnesota*

The significance of individual change has a wide range of application in educational or psychological testing. This study further evaluated two recent hypothesis testing methods, the Z statistic and LR statistic, in Finkelman et al. (2010) to determine the significance of individual change from two measurement occasions. A new test statistic, the LM statistic, for detecting individual change was also introduced. The performance of the three hypothesis testing methods was evaluated in the context of both conventional tests (CT) and adaptive measurement of change (AMC) based on observed Type I error rates and power. Using Monte Carlo simulation, the effect of item discriminations (low, medium and high), bank information shape (peaked and flat), and test length (15, 30, and 50 items) on the performance of the test statistics were also investigated. The two likelihood-based methods, the LR and LM statistics, displayed a better balance of Type I error rates and power than the Z statistic under most conditions. The LM statistic improved on the two existing methods in AMC with better adherence to Type I error rates and better power. AMC was more effective and efficient in detecting significance of change than CTs.

Mean Structure Analysis of MTMM Data

Poster Presentation

Mina M. Lee, *Sungkyunkwan University*
Soonmook Lee, *Sungkyunkwan University*

Covariance structure analysis of Multitrait Multimethod (MTMM) data has been around for a long while. However, mean structure analysis has not been attempted on MTMM data because it has not been viewed meaningful to impose a mean structure on a single set of data such as MTMM data. However, due to different methods used in collecting MTMM data it is reasonable to treat MTMM data as a special type of multi-group data. In line with that idea configuration invariance and metric invariance have been tested with MTMM data. However, it has not been attempted to test for scalar invariance and factor mean comparison with MTMM data. Difficulty of obtaining meaningful factor means and intercepts of observed variables has been the hurdle in attempting substantive interpretation of a mean structure for a single set of data such as MTMM data. We can resolve this problem by introducing an effect coding approach to estimating intercepts of observed variables. Then factor means are estimated as a linear combination of observed variable means. Consequently factor means can be compared with each other as well as intercepts can be compared across methods.

Test Assembly Implications for Providing Reliable and Valid Subscores

Paper Presentation

Minji Lee, *University of Massachusetts Amherst*
Kevin Sweeney, *College Board*
Gerald Melican, *College Board*

This study explores the relationship between factor correlation and the reliability of subscore estimates, providing a guideline with respect to psychometric properties of useful subscores. In addition, it compares subscore estimation methods with respect to reliability, distinctness, and accuracy. The subscore estimation methods used in the current study include Haberman augmentation (Haberman, 2008) and multidimensional item response theory (MIRT) estimation. The study shows that there is no estimate that perfectly meets all three criteria: Augmented subscores and MIRT estimates tend to be less distinct, and the observed subscores tend to be less reliable and less accurate.

Ability Purification to Reduce Type I Error Rates of the Detection of Answer Copying

Paper Presentation

Seo Young Lee, *University of Wisconsin-Madison*

In the detection of answer copying, it is important to reduce type I error rates because detecting honest students as copying may cause serious problems. The accurate estimation of ability parameters would lead to better performance in the copying detection. However, it is not surprising that when answer copying on tests occurs, the estimation of ability containing copied items would be inaccurate. In this study, a procedure was suggested to improve ability estimation so that the type I error rates of copying detection would be reduced. Ability parameters were estimated after item responses suspected as copying were excluded. The detection of copying using was performed with the purified ability parameter estimates. Through simulation studies, this approach showed that the bias of ability estimates and the type I error rates of answer copying detection were reduced.

Multidimensional Extension of Multiple Indicator Multiple Cause Models to Detect DIF

Poster Presentation

Soo Youn Lee, *Rutgers University*
Youngsuk Suh, *Rutgers University*
Okan Bulut, *American Institutes for Research*

Studying differential item functioning (DIF) continues to shed light on both applications and methodologies. A number of studies have generally found multiple indicator multiple cause (MIMIC) models to be an effective tool in DIF detection for individual items (Finch, 2005; Shih & Wang, 2009; Wang & Shih, 2010; Woods, 2009). Currently Lord's Wald test is considered superior to the MIMIC models and the generalized Mantel-Haenszel method in detecting DIF (Woods et al., 2013). The use of MIMIC models based on confirmatory factor analysis for detecting uniform DIF has been extensively studied in extant literature. Recently, identifying nonuniform DIF with MIMIC-interaction models has been studied in the unidimensional item response theory (IRT) framework (Woods & Grimm, 2011). Two methods for detecting both uniform and nonuniform DIF will be investigated: MIMIC models and the improved version of Lord's Wald test (Woods et al., 2013). The goal of the current study is to extend the MIMIC models in the context of multidimensional IRT and to compare this new approach with the Lord's Wald test using a Monte Carlo study. Simulation factors include sample size, DIF magnitudes, and ability distribution difference.

Comparison of Parameter Estimation Methods for the Logistic Positive Exponent Model

Poster Presentation

Sora Lee, *University of Wisconsin, Madison*
Daniel M. Bolt, *University of Wisconsin, Madison*

Samejima (2005) proposed logistic positive exponent (LPE) models for dichotomously scored test items, which introduce an asymmetry to the item characteristic curves that can represent variability in item complexity. She also proposed a possible strategy to estimating the exponent parameter using nonparametrically estimated ICCs (Samejima, 1995). We compared an implementation of this approach and a fully Bayesian approach recently proposed by Bolfarine and Bazan (2010) using simulated and real item response data.

An Alternative Diagnostic Measure for Detecting Nonlinear Relationships between Latent Variables

Paper Presentation

Taehun Lee, *University of Oklahoma*
Li Cai, *University of California, Los Angeles*

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

Psychometric Properties of Mixed-Format Tests

Paper Presentation

Won-Chan Lee, *University of Iowa*
Jiwon Choi, *University of Iowa*
Yujin Kang, *University of Iowa*
Stella Kim, *University of Iowa*

Psychometric properties of test scores often are evaluated to produce results that can be used to support and inform the use and interpretation of the scores (Kolen & Lee, 2011). This study considers psychometric properties of scale scores on mixed-format tests that consist of a mixture of multiple-choice (MC) and constructed-response (CR) items. Scale scores on various mixed-format tests are evaluated with respect to conditional standard errors of measurement, reliability, classification consistency and accuracy, and equity properties of equating. Both unidimensional and multidimensional item response theory (MIRT) frameworks are used and compared for assessing psychometric properties of the scores on the mixed-format tests. More specifically, unidimensional, bi-factor, simple structure, and full MIRT versions of the three parameter logistic and graded response model combinations are considered. The primary purposes of this study are to (1) develop procedures for assessing aforementioned psychometric properties based on various unidimensional and MIRT models, and (2) apply the procedures to a large number of real mixed-format tests that differ in subject areas, test length, composition of MC and CR items, and structure of multidimensionality.

Multivariate Maximum Entropy and Minimum Cross Entropy Distributions with Skewness, Kurtosis and Polychoric Correlation

Poster Presentation

Yen Lee, *University of Wisconsin, Madison*
David Kaplan, *University of Wisconsin, Madison*

When conducting robustness research the focus of attention is typically on the normality assumption with skewness and kurtosis often used to measure the non-normal level of the data. Research is often conducted via Monte Carlo methods for simulating distributions with skewness and kurtosis constraints. The maximum entropy (MaxEnt) distribution and minimum cross entropy (MinxEnt) with skewness and kurtosis constraints are presented here as a method to conduct Monte Carlo research. Subject to skewness and kurtosis constraints, the MaxEnt distribution possesses the most uncertainty and the MinxEnt distribution would be the one "nearest" to the distribution researcher specified previously. Using these distributions, researchers could investigate robustness to variations in the shape of distributions with the same level of non-normality. A procedure is presented here to estimate the MaxEnt/MinxEnt distribution that satisfy the constraints and to generate the corresponding data. The program is written in R. The procedure to generate the proposed distributions uses Lagrange multipliers, sequential least-squares quadratic programming and particle swarm optimization. The procedure is evaluated by checking the correctness of the parameters and the entropy of the estimated distributions. In addition, the parameter recovery rate and the moment distributions of the generated data are also estimated to understand the properties of the estimated distributions. Implication for research on the robustness of factor analysis to non-normality is provided.

Evaluating the Nonlinear Random-Effects Mixture Model for Repeated Measures Under Conditions that May Result in Spurious Classes

Poster Presentation

Houston F. Lester, *University of Nebraska-Lincoln*
Chaorong Wu, *University of Nebraska-Lincoln*
Natalie A. Koziol, *University of Nebraska-Lincoln*

Codd and Cudeck (2014) presented a general model (i.e., nonlinear random-effects mixture models for repeated measures) that can include different measurement schedules, data that are missing at random, as well as nonlinear functions of random-effects, predictors, and residuals. This simulation study aims to evaluate this model under conditions that have been found to result in spurious classes in structural equation model mixtures (Bauer & Curran, 2004) as well as conditions that are unique to the model of interest. To investigate the effects of these conditions, data will be generated from a two class model where one class follows the Richards growth curve model (i.e., $Y=\alpha[1+(\delta-1)\ e^{(-\kappa(X-\tau_i)}]^{(1/1-\delta)}; \delta\neq1$) and the other class has a linear trend. The mixing proportions for these two classes will be .6 and .4, respectively. The conditions of interest are: model misspecification/reparameterization of the generation model (i.e., the logistic growth model and a reparameterization of the Richards model, $Y=\alpha(1+b_i e^{-\kappa X})^M$, degree of class separation (i.e., determined by the covariance matrix difference between the two classes), and estimation method (i.e., maximum likelihood and Bayesian estimation with and without informative priors). These conditions will be evaluated in terms of the effect that they have on the number of classes selected.

Generalized Sequential Probability Ratio Test for Separate Families of Hypotheses

Paper Presentation

Xiaoou Li, *Columbia University*

We generalize the seminal work of the sequential probability ratio test to the problem of composite null hypothesis against composite alternative hypothesis that are completely separated from each other. In particular, the generalized likelihood ratio statistic is considered and the stopping rule is the first boundary crossing of the generalized likelihood ratios statistic. We show that this sequential test is asymptotically optimal without assuming distributions of exponential families. The current work has applications in psychometrics. For instance, the underlying distribution for null hypothesis could be a normal ogive model that is not of the canonical form of exponential family and the three-parameter logistic model that includes a guessing parameter.

On Closeness Between Factor Analysis and Principal Component Analysis under High-dimensional Conditions

Poster Presentation

Lu Liang, *University of Hawaii at Manoa*
Kentaro Hayashi, *University of Hawaii at Manoa*

It has been known that factor analysis (FA) and principal component analysis (PCA) lead to an identical loading matrix as the number of observed variables (p) increases without limit while the ratio of the number of factors (m) to p goes to zero. Because the sample covariance matrix S is properly defined only if the sample size (N) is larger than p, it implies that N also increases without limit. A recent theoretical study finds an extra condition that the ratio of 'square root of p' to N should go to zero. We examine the extra condition with an extensive simulation study. Furthermore, we also examine how the formulas for asymptotic standard errors derived under finite N and p and the formulas derived under increasing N and p without limit are connected for both FA and PCA.

IRT Model Selection for a Passage-Based Test Using Model-Fit Criteria

Paper Presentation

Euijin Lim, *University of Iowa*
Kyung Yong Kim, *University of Iowa*
Shichao Wang, *University of Iowa*
Won-Chan Lee, *University of Iowa*

Choosing an appropriate IRT model is not a straightforward process. The characteristics of items, dimensionality of the test, model-data fit, and item fit can be informative for decision making. When a test consists of questions related to a passage, a researcher has several alternatives: applying unidimensional dichotomous models (e.g., two or three parameter logistic model) ignoring passages; using unidimensional polytomous models (e.g., graded response or generalized partial credit model) treating items related to a passage as a single item with multiple score categories; or fitting multidimensional models (e.g., bi-factor or testlet response model) considering passage effects. Comparing

various possible models using model-fit criteria could provide important information about which model best explains the test data from a passage-based test. This study focuses on applying various model-fit criteria to a passage-based tests to determine the most appropriate model. The consistency of results across different model-fit criteria will also be examined. Two sets of large-scale test data are used and eight different IRT models are compared with respect to Akaike's information criterion (AIC), the Bayesian information criterion (BIC), the cross-validation log-likelihood (CVLL), and the Orlando and Thissen (2000) S-X2 item fit chi-square statistics.

The Effect of Artificially Compromised Items for Computerized Adaptive Testing

Poster Presentation

EunYoung Lim, *Korea Institute for Curriculum and Evaluation*

The purpose of this research is to explore the effect of artificially compromised items for computerized adaptive testing. It is obvious that compromised items will lead to overestimation of examinee ability, but there are not much empirical evidence or research to show our intuition. For this study, I performed a computerized adaptive testing simulation study to explore the effect of artificially compromised items with the maximum information item selection method (MI), α-stratified multi stage method (AS), and Sympson-Hetter method incorporated with these two item selection methods (MISH, ASSH). Change rates of compromised items from a wrong answer to right answer have been set as 10%~100%. Along with the change rate, bias and MSE for measurement precision and for item bank usage were compared. When an item bank did not have compromised items, MI showed the lowest bias and MSE. However, as the change rates had been increased, MI's bias and MSE dramatically increased. Also biases and MSE's of other algorithms dramatically increased, but ASSH and AS show lower bias and MSE than MI and MISH.

Accuracy of Diagnostic Classification under Various Conditions in the Rule Space Method

Paper Presentation

YeongYu Lim, *Georgia Institute of Technology*
Susan E. Embretson, *Georgia Institute of Technology*

The rule space method (RSM; Tatsuoka, 1983; 1985, 2009) is a famous cognitive diagnostic method based on the item response theory (IRT) models. In the present study, a customized algorithm with SAS macro was developed for the RSM analysis. A simulation study will be conducted to evaluate (1) accuracy of diagnostic classification in the RSM and (2) the impact of various factors on the consistency of the diagnostic results. The simulation variables include the number of attributes (6, 8, and 9), cut points (p-value = .50, .60, and .70), attribute correlations (low, medium, and high). The number of items is 60 for the three types of tests, and three Q-matrices were constructed for 6, 8, and 9 attribute conditions. The number of examinees is 1,000 for each condition and the number of replication of the simulations is 20. The simulation data will be generated with SAS macro. The RSM will be applied to the response data, and finally, individual attribute mastery probabilities will be obtained as the classification results. Correct classification rate (CCR), average Signed biases (ASB), and root mean square error (RMSE) will be computed to discuss accuracy of diagnostic classification across various conditions.

Parameter Estimation of Cognitive Diagnosis Models

Paper Presentation

Youn Seon Lim, *University of Illinois at Urbana-Champaign*

In this study I propose a pure simulation-based approach for computing joint maximum likelihood estimates in cognitive diagnosis models, carried out by means of Markov Chain Monte Carlo (MCMC) methods. The central theme of the approach is to reduce the complexity of models to focus on their most critical elements to estimate. In particular, an approach analogous to joint maximum likelihood estimation is taken, and the latent attribute vectors are regarded as structural parameters, not parameters to be integrated out of the problem. When doing so the joint distribution of the latent attributes does not have to be specified, which reduces the number of parameters in the model. By implementing the MCMC algorithm that simultaneously evaluates and optimizes the likelihood function without resorting to a gradient method, the estimates are obtained. This streamlined approach performs as well as more traditional methods for models such as the DINA, and affords the opportunity to fit more complicated models in which other methods may not be feasible.

Could Structure and Measurement Parts of an SEM Model be Assessed Using the Same Criteria?

Poster Presentation

Ching Lin, *National Cheng Kung University*
Chung-Ping Cheng, *National Cheng Kung University*

Structural equation modeling has become a popular analytic method in social and behavioral sciences. An SEM model can be assessed by many indices such as ML statistic and fit indices. Hu and Bentler (1998) have reported that different fit indices may be sensitive to different parts (i.e., structure and measurement parts) of a model, respectively. To our knowledge, there is little research following this suggestion. The current study will investigate this problem by controlling the parameter values of the two paths on the same statistical power, and several conditions including nonnormality, and degree of model imperfection will be manipulated. The performance of ML statistics and fit indices such as CFI, NFI, SRMR, and RMSEA will be recorded, and the performance of the structural and measurement parts will be compared.

Using Bayesian Inference Networks to Model Student Practices in NAEP Science Interactive Computer Tasks

Paper Presentation

Johnny Lin, *Educational Testing Service*
Yue (Helena) Jia, *Educational Testing Service*
Margarita Olivera-Aguilar, *Educational Testing Service*
Yoav Bergner, *Educational Testing Service*

The National Assessment of Educational Progress (NAEP) science interactive computer tasks (ICTs) assess student science knowledge and practices in a computer based environment by simulating natural or laboratory settings. Traditional Rasch-based unidimensional item

response theory (IRT) models are limited in scoring ICTs especially when the researcher is interested in modeling complex structural relationships among latent traits. Alternatively, Bayesian Inference Networks (BINs) allow more complex specification of both the proficiency and evidence models as part of the Evidence Centered Design paradigm (Mislevy, Steinberg & Almond, 2003). This presentation will draw connections between BINs and IRT, coat-tailing previous research that imposes an IRT functional form within a BIN framework, using reconfigured task parameters and discretized latent ability (Almond et al., 2007). The second part of the talk will demonstrate how IRT-BINs may be used to model student competencies in the 2009 NAEP science ICTs, and in particular how this method may account for conjunctive and compensatory relationships between students' prior knowledge about science and task-relevant context effects.

When the Predictions Encounter the Observations: What Does it Tell? - Exploring the Existence of Marketing Constructs from a Psychometric Perspective

Poster Presentation

Shan Lin, *Norwegian School of Economics*

Due to the limitations of theories and statistical techniques, traditional measurement towards latent variables is not enough to tell whether the construct of interest can represent some sort of reality. Additionally, the traditional way does not address the problem directly. This proposal aims to call for more attention towards the latent variables and addressing measurement from some new perspectives.

Successive Blocks Combined with Item Pocket Method to Allow Item Review in Computerized Adaptive Testing

Poster Presentation

Zhe Lin, *Beijing Normal University*
Tao Xin, *Beijing Normal University*
Ping Chen, *Beijing Normal University*

Most computerized adaptive tests (CAT) do not allow examinees to review items due to serious deterioration of the measurement efficiency and extra cheating strategies. It has been verified that the above problems will be significantly diminish by setting successive blocks or an item pocket in CAT's process. However, both methods have their own limitations. The former does not allow examinees to skip items. And it has to set too many successive blocks that will increase adverse effecting due to frequent decision making about going to next block. The latter needs to set a proper IP size and may not avoid Wainer cheating strategy if a large size was provided.   This study proposes a combined method, setting fewer but larger blocks containing a smaller size IP in each, which could overcome those disadvantages and maintain their benefits. A series of simulations were conducted to compare the combined method with the IP method, especially the robustness of those methods against cheating strategies. The results suggested that the combined method was more effective in measurement efficiency and avoidance of cheating strategy. As number of blocks increases, both MAE and RMSE have a slight decrease, and are more close to the non-review condition.

Mediation Analysis using Multimethod Designs with Structurally Different and Interchangeable Methods: An Application in Personality Psychology

Paper Presentation

Kaylee Litson, *Utah State University*
Lesther Papa, *Utah State University*
Christian Geiser, *Utah State University*
Ginger Lockhart, *Utah State University*
Michael Eid, *Freie Universiteit Berlin*

Researchers are often interested in how an exogenous variable X is indirectly related to an outcome variable Y via a mediator M. Frequently, such indirect associations are examined with modern path-analytic methods of statistical mediation analysis (e.g., MacKinnon, 2008). In addition, researchers frequently use multitrait-multimethod (MTMM) measurement designs (Campbell & Fiske, 1959) in which X, M, and Y each are assessed with multiple methods (e.g., multiple reporters). The use of MTMM designs enables researchers to study the convergent and discriminant validity of different methods. We present a statistical model that combines the advantages of modern mediation analysis with the advantages of modern confirmatory factor analysis approaches to MTMM analysis. Our modeling framework allows studying mediation with structurally different methods, interchangeable methods, and a combination of structurally different and interchangeable methods. We present an application of the model in which mood regulation is assessed as a potential mediator between personality traits and actual mood. In this example, N = 482 individuals were assessed by two friends (interchangeable method) and by themselves (structurally different method).

Searching for Alternative Methods for Propensity Score Estimation

Poster Presentation

Dong Liu, *Renmin University of China*

Logistic regression (LR) has been proven to be a coarse method for the estimation of propensity score, especially in nonlinear and additive cases. Alternatives of LR, therefore, have been proposed in recent years. However, almost none of the proposed alternatives perform consistently better than LR. The aim of this study is to search for such a consistent alternative to replace LR. The alternative PS estimation methods include: Naive Bayesian, Random Forest, GUIDE Forest, and Logistic Model Trees. Performance outcomes include absolute bias; mean squared error; and standard error of the treatment effect estimate and covariate balance. In this study, the performances of the alternatives relative to LR are examined via simulation. The study used the same data generation setup and selection models (A-G) as Setoguchi et al.(2008). The results showed that Naive Bayesian performed consistently better than LR and all other alternatives even in linear cases. In the case with nonlinearity and interaction terms, the absolute bias averaged over 1,000 replications was 20% for correctly specified LR and less than 5% for Naive Bayesian.

Separating-strategy Priors for Covariance Matrices

Paper Presentation

Haiyan Liu, *University of Notre Dame*

In Bayesian modeling, the inverse-Wishart distribution is the most commonly used prior for an unknown covariance matrix. The Inverse-Wishart prior is proper and conjugate in a multivariate normal model. However, it has the same problems as the inverse-Gamma prior for a scalar variance. Although some other priors have been proposed, they are rarely used due to the limited analytical results that can be drawn. In this study, we consider a new type of separating-strategy prior. We decompose the covariance matrix into scalar variances and correlation coefficients. Priors are studied for each scalar variance and correlation coefficients. This flexible method enables us to utilize a large variety of priors for the covariance matrix. Through both simulation study and real data analysis, we compare the inverse-Wishart prior with the separating-strategy priors for both 2 by 2 and 3 by 3 covariance matrices. We find that the separating-strategy priors work better than the inverse-Wishart prior. The proper use of the separating-strategy priors is discussed.

Application of Nested Logit Model in Computerized Adaptive Testing

Paper Presentation

Tour Liu, *Beijing Normal University*
Tao Xin, *Beijing Normal University*

Multiple-choice items are very popular in many psychological and educational tests. However, these multiple-choice data are usually fitted with dichotomous item response models, like the 2-parameter logistic model (2PLM) and 3-parameter logistic model (3PLM). In this case, a large amount of information contained in distractors is ignored. Nested logit models (NLMs) proposed by Suh and Bolt (2010) were proved to be appropriate for multiple-choice data and could provide more information. In this research, the NLMs were used in computerized adaptive testing composed of multiple-choice items. The main purpose of this research is to detect whether the distractor information is helpful to enhance the effectiveness of computerized adaptive testing (CAT). Two conditions were taken into consideration. First, examinees were response normally. Some high ability examinees chose the distractors at the beginning of the CAT in the second condition. Three maximum Fisher information methods were used as an item selection procedure, the item Fisher information, the correct option Fisher information and the distractors Fisher information. The short CAT (10 items) and long CAT (20 items) were administrated. Compared to a dichotomous item response model (IRT), the recovery of ability parameters was considered.

Generalized Fiducial Inference for Binary Logistic Item Response Models

Paper Presentation

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

Generalized fiducial inference (GFI) has been proposed as an alternative to likelihood-based and Bayesian inference in mainstream statistics. Point estimates and confidence regions can be constructed from a fiducial distribution on the parameter space in a fashion similar to those used with a Bayesian posterior distribution. However, no prior distribution needs to be specified, which renders GFI more suitable when no a priori information about model parameters is available. In the current paper, we apply GFI to a family of binary logistic item response theory models, which includes the two-parameter logistic (2PL), bifactor and exploratory item factor models as special cases. The resulting fiducial distribution for item parameters satisfies a Bernstein-von Mises theorem, which justifies the asymptotic frequentist properties of the associated point and interval estimators. Random draws from the fiducial distribution can be obtained by the proposed Markov Chain Monte Carlo sampling algorithm. We investigated the comparative performance between GFI and maximum likelihood in parameter recovery using simulated 2PL data. The use of GFI in high-dimensional exploratory item factor analysis was illustrated by the analysis of a set of the Eysenck Personality Questionnaire data.

Limited Information Goodness-of-fit Statistics for Cognitive Diagnosis Models

Paper Presentation

Yanlou Liu, *Beijing Normal University*
Wei Tian, *Beijing Normal University*
Tao Xin, *Beijing Normal University*

The appropriateness of the Cognitive Diagnosis Model (CDM) to the response data needs to be evaluated in support of interpreting respondents' performance and test development. However, the so-called CDM-model-data fit is rarely considered. The most commonly used full information goodness-of-fit statistics $\chi^2$and G^2 in CDMs is out of use when the number of items is large and respondents are not enough (sparse contingency table). As a matter of fact, some kind of alternative methods have been proposed for this problem in the context of item response theory (IRT), some of which has also been extended to solve the CDM's sparse contingency table. Unfortunately, they were time-consuming and were not computationally feasible for practical applications. In this study, we extend limited-information goodness-of-fit statistics proposed by Cai et al. (2006) to the case of CDM to address the CDM model-data fit problem. Two simulation studies and one practical study were conducted. The first study is designed to show that χ^2, $G^2$and limited-information statistics provided proper empirical type I error rates for non-sparse contingency table. The second study aims to show that a limited-information statistic was the only proper method when contingency tables become sparse. The third study will provide practical evidence.

Time Series Models for Analyzing Development Within Twin Study Designs

Paper Presentation

Lawrence L. Lo, *Pennsylvania State University*
Peter C.M. Molenaar, *Pennsylvania State University*

A series of models are presented for analysis of twin time series data for behavior genetics research. The models are extensions of the idiographic filter ACE (iFACE) based on the extended Kalman filter and incorporate time-varying parameters. The presentation outlines basics of Kalman filtering for the purposes of model demonstration, and shows how an iFACE model set as a structural equation model (SEM) can be fit within a Kalman filter. All models are fit via maximum-likelihood procedures using the Kalman filter prediction error decomposition. The first set of models utilizes time-varying components of the iFACE model to allow for developmental changes in parameters such as heritability or shared/non-shared environmental influence. These are essentially time-varying factor loadings within a SEM structured iFACE model. The final model allows for time-varying genetic correlation where the genetic relatedness of siblings is allowed to change across development. This model utilizes a time-varying structural correlation for achieving this purpose. A simulation study demonstrates the capability of the models to capture these types of developmental changes with simulation conditions set to several possible alternative developmental hypotheses. Several possible directions for future research based on these modeling procedures are proposed and outlined.

What Lies Beneath Situational Judgement Tests? A Psychometric Evaluation Using Multidimensional Item Response Theory

Poster Presentation

Aiden Loe, *University of Cambridge*

Situational judgement tests (SJTs) are personnel selection tests. Several theories have been used to describe what SJTs measure. These theories, however, are inadequate in providing a universal framework. Factor analysis is the most prevalent analytical method to model construct-based SJTs. However, results from factor analysis often reveal low internal consistencies because SJTs are multidimensional. This suggests a broad variety of factors contributing to the internal structure of SJT, making it difficult to elucidate and validate the constructs via factor analytical techniques. On the contrary, multidimensional item response theory (MIRT) holds a promising approach to validate SJT constructs. MIRT can account for the multidimensionality of items far better than factor analysis. Furthermore, MIRT models can address the testlet effects. Testlet effects are a separate group of factors that influence the construct validity. MIRT models can filter these testlet effects when modelling SJTs. The present study will use the multidimensional generalised partial credit (MGPC) model and the MIRT bi-factor model to recover latent dimensions from a SJT data of 3128 graduate participants. The recovered latent dimensions will be correlated with cognitive ability scores and performance data to measure for criterion-related validity. The findings and results will be presented at the conference.

Psychometrics of Assessment in College STEM Classes

Paper Presentation

Eric Loken, *Pennsylvania State University*
Lawrence Lo, *Pennsylvania State University*

Although there has been recent interest in evidence-based approaches to improve college STEM teaching, there has been little research on the state of college testing. At many large universities, introductory classes can enroll thousands of students. Universities are effectively acting like major testing organizations, even though very little attention is paid to the quality of assessment. We explore what it might mean to take a more research based approach to the construction, administration, and interpretation of college level tests. Examining data from tens of thousands of tests in college level statistics, physics, chemistry, and astronomy, we consider the fit of various psychometric models as well as criteria of test fairness (equating and differential item functioning). We also look at pre and post instruments used to evaluate learning at the class level. Attempting to characterize the student model, and in particular the model of change associated with successful learning is a challenge for one semester courses. We situate the current state of college testing relative to broader developments in principled assessment design, accountability, and new innovations in the delivery of college courses.

Automated Scoring Procedure Using Diagnostic Classification Models

Paper Presentation

Diane Losardo, *Amplify Education*
Margi Dubal, *Amplify Education*

In the Item Response Theory (IRT) literature, observations are often scored in a two-step procedure, in which first an IRT model is calibrated, then the estimated model is used to obtain continuous scores for subsequent samples. Such a procedure may be beneficial when using Diagnostic Classification Models (DCMs), as resulting scores would now be a probability of mastery for one or more skills. Roussos et al. (2007), developed software (Arpeggio) that implements this scoring procedure using the Fusion Model, which is a particular parameterization of a DCM. However, a general automated procedure subsuming all DCM structures is not readily available. In this talk we present a procedure that automates the process of scoring individuals using a DCM as estimated within the loglinear cognitive diagnostic modeling (LCDM; Henson, Templin, & Willse, 2009) framework that allows for the estimation of a host of DCMs. We describe a python plug-in that auto-scores observations using a Markov Chain Monte Carlo procedure and an R script that generates useful plots and diagnostic information. We then illustrate this process using Amplify Education assessment data by describing the calibration process of a DCM, illustrating the automated scoring using the Python plug-in, and generating plots using the R script.

Using the Jackknifing Method to Evaluate Anchor Stability with the NEAT Design

Paper Presentation

Ru Lu, *Educational Testing Service*
Shelby Haberman, *ETS*
Jinghua Liu, *SSAT*

In the non-equivalent groups with anchor test (NEAT) design, the anchor test is used to adjust form differences after accounting for the ability differences. Just as different equating samples could produce different equating results, different anchor tests could produce different equating results. This study proposes a measure of anchor stability to indicate the equating variability that is due to anchor item sampling. A jackknifing method is used in obtaining the anchor stability measure. Two real data examples from a large-scale testing program are evaluated. Three factors (ability difference, anchor test length, and equating methods) that could affect the measure of anchor stability are manipulated. The results show that anchor test length and equating methods have limited impact on the anchor stability measure. Ability difference has the largest impact on the anchor stability measure. The real data examples also show that the effect of anchor item sampling could be rather large when equatings are conducted with less satisfactory conditions.

Aggregating Time Series: Illustration Through an AR(1) Model

Paper Presentation

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

Intra-individual analysis has become popular as advanced by Nesselroade, Molenaar, & colleagues. Laws governing the inter-individual relationship (cross-sectional data) may not be applied to intra-individual (time series data) (e.g, Molenaar, 2004; Nesselroade & Ram, 2004). For intra-individual single time series analysis, many methods are available (e.g., Cattell, Cattell, & Rhymer, 1947; Nesselroade, et al., 2003). In practice, it is always tempting to collect data from more than one participant. However, there are relatively few methods available for dealing with the analysis of multiple time series (e.g., Cattell & Scheier, 1961; Molenaar, et al., 2003; Nesselroade & Molenaar, 1999). This study aims to illustrate issues related to pooling time series data through the simplest time series model, autoregressive model of order 1, AR(1). After the literature review, the AR(1) model is presented. Next, methods to pool time series data are introduced. We propose different methods to pool time series data, which include connecting data directly, pooling individual's likelihood functions (both pooling conditional likelihood and pooling exact likelihood), and multivariate time series analysis. The focus is on pooling individual's likelihood functions, presented by mathematical analysis and simulations studies. Pros and cons of each method are also compared. Finally, some practical implications are addressed.

From Tukey to Statistical Learning: Gaining New Insights from Big Data

Invited Speaker

Gitta Lubke, *University of Notre Dame*

The topic of "big data" has attracted so much attention recently that it makes you wonder whether this is just another hype. Large data sets are, however, a reality, and this presentation focuses on methods designed to explore and analyze large data sets in the absence of knowledge concerning the structural relations between the measured constructs. The emphasis is on gaining new insights rather than on confirmation, similar to the exploratory data analysis methods described by Tukey. The development of statistical learning algorithms has increased the feasibility of exploring extremely large data sets, including data where the number of observed variables is much larger than the number of subjects. The challenge is to integrate psychometric knowledge and methodology in the field of statistical learning.

Bayesian Model Comparison Using P-Values

Paper Presentation

Joseph F. Lucke, *State University of New York at Buffalo*

I introduce a procedure for Bayesian model comparison (BMC) using p-values. Bayesians are critical of using p-values for inference, but here I propose to use them as data. The BMC then uses the posterior probabilities of two models given the p-value. From Hung et al. (1997) and Donahue (1999), the distribution of the p-statistic is known. Following DeSantis (2004), I use the Bayes Factor to find the regions for the p-values that yield confirming evidence for each model and the region that yields weak evidence confirming neither model. From this information and the prior probabilities for each model, one can determine the probability at which an observed p-value provides confirming, weak, or misleading evidence for the models. Likewise, one can determine the sample size required to achieve a confirming Bayes Factor at a given level of probability. For reference, I also compare this approach to standard frequentist power analyses. The capability to conduct BMC for p-values allows the Bayesian to easily comment on the evidence provided by results based on p-values.

Polyserial versus Pearson Correlations in Confirmatory Factor Analysis with Continuous and Ordinal Variables

Paper Presentation

Hao Luo, *The University of Hong Kong*
Fan Yang-Wallentin, *Uppsala University*
Vivian HT So, *The University of Hong Kong*

Confirmatory Factor Analysis (CFA) is a widely adopted technique in social and behavioral sciences in testing whether measures of a construct are in line with researchers' hypothesis. In the process of fitting a CFA model, ordinal variables are commonly encountered in many empirical investigations. Practitioners usually treat ordinal data as continuous and fit the model using maximum likelihood ignoring the inherent non-normal property. More cautiously, some researchers use asymptotically distribution free methods represented by weighted least squares (WLS) and diagonally weight least squares (DWLS). To adjust for the situation of having mixed data with both continuous and ordinal variables, a better approach is to use polyserial correlations and fit the models using methods such as unweighted least squares (ULS), WLS, and DWLS. In this current simulation study, we compare the performances of these estimators in combination with both polyserial and Pearson correlation coefficients. Monte Carlo experimental conditions include different sample sizes, distributional settings, number of categories and the number of ordinal variables. Comparison criteria are set to be parameter estimates, their standard errors and common chi-square measures of fit. The advantages and disadvantages of using polyserial correlation instead of Pearson correlation are discussed in order to provide guidance to practitioners.

Effect of Parental Warmth on Children's Sociability Moderated Mediating Effect

Poster Presentation

Yachen Luo, *Beijing Normal University*

In this study, we focus on the mechanism of how parental warmth may affect children's sociability. This study constructed a moderated mediation model to examine whether perceived parental warmth mediated the relation between parental warmth and children's sociability, and whether this mediating process was moderated by gender. 2942 junior high school children and their parents from Beijing in the People's Republic of China participated in this study and completed questionnaire measures of parental warmth, children's perceived parental warmth and adolescents' sociability. The results indicated that (1) Children's perceived maternal warmth is obviously higher than perceived paternal warmth. (2) Perceived parental warmth played a mediating role in the relationship between parental warmth and children's sociability; (3) Gender moderated the mediated path through perceived parental warmth, such that this indirect effect was much stronger for junior high school girls relative to junior high school boys. Thus, both mediating and moderating effects exist in the association between perceived parental warmth and children's sociability.

An Empirical Study of the Impact of the Choice of Persistence Model upon Teacher Effect Estimates in Value Added Modeling

Paper Presentation

Yong Luo, *American Nurses Credentialing Center*
Hong Jiao, *University of Maryland, College Park*
Robert Lissitz, *University of Maryland, College Park*

The generalized persistence (GP) model developed by Mariano, McCaffrey, and Lockwood is a general multivariate model for teacher effect estimation in value added modeling (VAM) when construct shift exists across grades. It is a generalized case of other persistence models such as the zero persistence (ZP) model, the complete persistence (CP) model, and the variable persistence (VP) model. Other than the empirical example used in the original paper by the authors, no other empirical studies that apply the GP model to real data are found in literature despite the need for more empirical studies to investigate the performance of the GP model in comparison to other persistence models, and it remains unclear how the choice of different persistence model impacts teacher effect estimates. Using longitudinal data from an Eastern state from 2008 to 2010 that include four cohorts and two subjects (English and Mathematics), the current study investigates whether the choice of different persistence model results in different teacher effect estimates and whether the pattern remains invariant between English and Mathematics. In addition, the findings will be compared with those in the original paper to investigate their generalizability.

A Multidimensional Nonparametric Estimator of a Monotone in Each Latent Trait Item Characteristic Curve

Paper Presentation

Mario Luzardo, *Universidad de la República*

It is very important develop models for monotone items in IRT. The smoothed nonparametric model is not necessarily monotone, so Ramsay (1998) introduced a procedure to obtain isotone ICCs, but his method is semiparametric because the ICC requires that a second-order differential equation be satisfied, which makes the procedure generally inconsistent. Recently, Lee (2002, 2004 and 2007) proposed the use of an isotonic regression method promoted by Barlow, Bartholomew, Bremmer, and Brunk (1972) as well as Robertson, Wright and Dykstra (1988), a least squares method for data fitting under order restrictions. Luzardo and Forteza (2013) presented a unidimensional model based on nonparametric isotone regression, estimating monotonic functions from non monotonic estimators. Now I introduce a multidimensional nonparametric estimator of monotone in each latent trait item characteristic curve. This presentation shows a five step algorithm easy to implement for the two dimensional case and compares its performance with the parametric IRT models.

Detection of Differential Item Functioning Using the Lasso Approach

Paper Presentation

David Magis, *KU Leuven*
Francis Tuerlinckx, *KU Leuven, Belgium*
Paul De Boeck, *Ohio State University*

The purpose of this talk is to present a novel approach to detect differential item functioning (DIF) among dichotomously scored items. Unlike standard DIF methods that perform an item-by-item analysis, we consider a logistic regression model including item-group interaction (i.e. DIF) effects of all items simultaneously. The method is based on penalized maximum likelihood estimation of a model with a lasso penalty on all possible DIF parameters. Optimal penalty parameter selection is investigated through several known information criteria (such as AIC and BIC) as well as a newly developed weighted alternative. A simulation study was conducted to compare the global performance of the suggested 'lasso DIF' method to the logistic regression and Mantel-Haenszel methods, and to evaluate the different optimal penalty parameter selection methods. It is concluded that for small samples the lasso DIF approach globally outperforms the logistic regression method, and also the Mantel-Haenszel method, especially in the presence of item impact, while it yields similar results with larger samples.

Modeling and Testing Differential Item Functioning in Unidimensional Binary Item Response Models with a Single Continuous Covariate: A Functional Data Analysis Approach

Paper Presentation

Brooke Magnus, *University of North Carolina at Chapel Hill*
Yang Liu, *University of North Carolina at Chapel Hill*
David Thissen, *University of North Carolina at Chapel Hill*

Differential item functioning (DIF; see Angoff, 1993, for a review), referring to differences in item characteristics across groups above and beyond the group-level differences in the latent variable of interest, has long been regarded as an important item-level diagnostic. The presence of DIF impairs the fit of the single-group item response model being used, and calls for either model modification or item deletion in practice, depending on the mode of analysis. Methods for testing DIF with continuous covariates, rather than categorical grouping variables, have been developed (e.g., Bauer & Hussong, 2009; Skrondal & Rabe-Hesketh, 2004); however, they are restrictive in parametric forms, and thus are not ideal for describing involved interactions among latent variables and covariates. In the current study, we formulate the probability of endorsing each test item as a general bivariate function of the unidimensional latent trait and the single covariate, which is then approximated by a two-dimensional smoothing spline (Currie, Durban, & Eilers, 2006). The accuracy and precision of the proposed procedure is evaluated via Monte Carlo simulation. In cases where anchor items are available, we propose an extended model that simultaneously estimates item characteristic curves (ICCs) for anchor items, ICCs conditional on the covariate for non-anchor items, and the latent variable density conditional on the covariate - all using regression splines. A permutation DIF test is developed and illustrated with a simulated data example.

Quantities, Quantification, and the Necessary and Sufficient Conditions for Measurement

Paper Presentation

Luca Mari, *School of Industrial Engineering, Universita Carlo Cattaneo-LIUC, Castellanza (VA), Italy*
Andrew Maul, *University of Colorado, Boulder*
David Torres Irribarra, *University of California, Berkeley*
Mark Wilson, *University of California, Berkeley*

It is remarkably difficult to provide a fully satisfactory definition of the concept of measurement. In part this is due to the fact that measurement is a multifarious, dynamic, and designed-on-purpose human activity. To respect this fact, a definition of measurement cannot be so narrowly construed as to apply to only one area of scientific activity (e.g., physics); on the other hand, the definition cannot be so permissive as to trivialize the concept to the point that measurement is not recognizably superior to, e.g., guesses or subjective judgments. One issue at the heart of this tension is the relationship between measurement, quantities and quantification. In particular, it is sometimes argued or assumed that (a) quantification is necessary for measurement, or (b) quantification is sufficient for (or even synonymous with) measurement. To assess the validity of these positions, the concepts of measurement, quantity and quantification should be independently defined and their relationships analyzed. We contend that the defining characteristics of measurement are not located in the structure of its inputs or outputs, but in the structure of the process and possibly in the quality of its results. Under this perspective, quantification is neither sufficient nor necessary for measurement."

Network Psychometrics: What All the Fuzz is About!

Paper Presentation

Gunter Maris, *CITO - University of Amsterdam*

The Ising network model from statistical mechanics can be represented as a marginal multi-dimensional IRT model. It is however a peculiar marginal IRT model, since its latent variable distribution depends directly on both the number of items, and on their statistical properties. Moreover, the statistical properties of items have to change when the number of items increases, in order for the model to not degenerate. When one considers learning from a classical psychometric point of view (ability increases through learning, all other things, like item properties, remain unchanged) this may seem weird. We argue that it is actually the classical psychometric perspective on learning and educational progress that is weird. If one wants to actually explain some of the well-established individual and group level learning phenomena, it is actually the peculiar marginal IRT model, that the Ising model is, that makes more sense. Using real and simulated examples we illustrate how learning can be conceived of as consisting of a process in which a network grows over time, with the structure changing, the magnetic field shifting, and the temperature dropping.

The Philosophical Foundations of Psychological Measurement

Paper Presentation

Andrew Maul, *University of Colorado, Boulder*
David Torres Irribarra, *University of California, Berkeley*
Mark Wilson, *University of California, Berkeley*

Measurement long played an important role in the physical sciences and natural philosophy. More recently, the psychological sciences have developed a variety of techniques that purport to be instances of measurement as well. However, it is not clear how the understanding of measurement invoked in psychological science applications accords with the understanding of measurement found in other scientific disciplines. Recently, some scholars have suggested that, by the criteria accepted in other disciplines, there may not yet have been any instances of successful measurement in the psychological sciences. Psychological scientists often react dismissively to such claims, and while we agree that philosophical debates are not a cause for despairing of activities that are successful by many important criteria, we also find that a sharper focus on conceptual clarity and coherence across the psychological and physical sciences has the potential to add a great deal to our efforts to improve such practices. In this paper, we argue that it is possible to formulate a philosophically coherent account of how measurement works in both the physical and the psychological sciences, and that such considerations also have practical value for psychometric practice.

Analysis of Item Wording Effects and Response Styles with Generalized Hierarchical IRT Models

Paper Presentation

Thorsten Meiser, *University of Mannheim*
Hanna Fleig, *University of Mannheim*

Positively worded items are commonly used together with reversed items to broaden the content domain of the item pool and to control for response styles (e.g., acquiescence). Previous research has shown, however, (a) that interindividual differences in reaction to reversed items may jeopardize the unidimensional trait structure and (b) that item reversal may affect the use of the rating format for individual items. While observed multidimensionality is usually addressed by structural equation models that include trait factors and method factors of item wording, effects of item reversal on the rating response format have rarely been considered. In this presentation, we propose multidimensional IRT models in the framework of generalized hierarchical models that allow researchers to analyze the role of response styles in item wording effects and to accommodate effects of item reversal on both the trait structure and the response format. Model specification is illustrated with empirical questionnaire data. Extraneous covariates and experimental manipulations of response styles are used to test implications for the interpretation of item wording effects. The results suggest that modelling item wording effects as residual effects (i.e., orthogonal to trait factors) may be inappropriate if method effects are assumed to include common response styles.

Testing Non-nested Structural Equation Models

Paper Presentation

Edgar C. Merkle, *University of Missouri*
Dongjun You, *University of Missouri*
Kristopher J. Preacher, *Vanderbilt University*

In this talk, we apply Vuong's (1989) non-nested likelihood ratio tests to the comparison of non-nested structural equation models. Similar tests have previously been applied in SEM contexts (especially to mixture models), though the non-standard output required to conduct the tests has limited their previous use and study. We review the theory underlying the tests and show how they can be used to construct interval estimates for differences in non-nested information criteria. Through both simulation and application, we then illustrate the tests' performance in non-mixture SEMs via the R package lavaan. The tests offer researchers a useful tool for non-nested SEM comparison, with barriers to test implementation now removed.

Fingerprint Bootstrapping: A Generic Method for Efficient Bootstrapping

Paper Presentation

Merijn Mestdagh, *KU Leuven*
Stijn Verdonck, *KU Leuven*
Denny Borsboom, *KU Leuven*
Francis Tuerlinckx, *KU Leuven*

Bootstrapping is a method for assessing the uncertainty of sample estimates (Efron & Tibshirani, 1994) and because of its versatility, the method enjoys an enormous popularity in various scientific disciplines. Unfortunately, bootstrapping can be a slow process because in each bootstrap iteration, an objective function (usually a likelihood) needs to be optimized. If this optimization problem is computationally intensive, the execution of the whole bootstrap process can become almost impossible in practice. We present an improvement of the bootstrapping process by increasing its speed versus accuracy ratio. The method capitalizes on the fact that subsequent bootstrap iterations all handle a similar optimization problem. The results of previous bootstrap iterations are therefore used to facilitate the optimization problems of following bootstraps. Each previous bootstrap data set can be characterized by a fingerprint (e.g., the gradient evaluated in the parameter estimate) and an optimum. Next, the relation between the fingerprint and the optimum is modeled non-parametrically. When this relation becomes more and more accurate, the fingerprints of following bootstraps will lead us closer and closer to their optimum, effectively bypassing the optimization problem. In this way, the speed of the bootstrap can be increased by a factor of ten (controlling for accuracy).

Multidimensional Item Factor Analysis with Semi-Nonparametric Latent Densities

Paper Presentation

Scott Monroe, *UCLA*
Li Cai, *UCLA*

In the item response theory (IRT) tradition, generalizations of models are often motivated by the substantive theory of researchers. Two important examples of this phenomenon are multidimensional IRT (MIRT), and the estimation of the latent variable distribution. In many ways, MIRT is based on the assumption that individuals often use multiple characteristics or skills when responding to test items. And, the belief that these characteristics may not be normally distributed has prompted research concerning estimation of the latent variable distribution. However, methodological and computational difficulties have, to some extent, hindered these generalizations. Capitalizing on recent methodological advances, this research proposes a multidimensional item factor model that uses a semi-nonparametric density (Gallant & Nychka, 1987; Wood & Lin, 2009) for the latent traits. Estimation is carried out via the Metropolis-Hastings Robbins-Monro (Cai, 2010) algorithm, which leads to maximum likelihood estimates as well as an estimate of the observed information matrix. A simulation study is conducted to evaluate the utility of the methodology, and empirical examples are provided.

The Rasch Model as a Special Case of A GLMM (Generalized Linear Mixed Model): Practical Advantages of This Unified Frame of Reference

Poster Presentation

Eiliana Montero-Rojas, *University of Costa Rica*

The Rasch model is a special case of an IRT model that has become very popular in recent years in applied educational and psychological measurement. This popularity is, due, in part, to certain particular properties that make it suitable for criterion referenced interpretations. It was precisely the Danish mathematician, statistician, and psychometrician Georg Rasch who uncovered these unique properties. More recently, it has been shown that the Rasch model is also a special case of a GLMM (Generalized Linear Mixed Model) where a logit is the linking function (for dichotomous items) and the fixed effects are the item difficulties. This unified approach offers some practical advantages such as the possibility of analyzing subject to subject variability by the estimation of the random effects and the inclusion of DIF analysis. Data from Costa Rica's 2012 application of PISA tests in Math are used to illustrate this new frame of reference and compare it with the traditional approach.

Multi-group Data and Item Non-response: A General Model Framework

Paper Presentation

Irini Moustaki, *London School of Economics and Political Science*
Myrsini Katsikatsou, *London School of Economics and Political Science*
Jouni Kuha, *London School of Economics and Political Science*

Sample surveys collect information on a number of variables for a randomly selected number of respondents. Among other things, the aim is often to measure some underlying trait(s) of the respondents through their responses to a set of questions. In the paper, we focus on cross-national surveys. The main research objective is to compare the distribution of the latent variables across countries (structural model). In some applications, latent variables will be considered continuous (e.g. ability) and in some other applications discrete (e.g. health state). Here, our focus will be primarily the modelling of item non-response and studying its effect on cross-countries comparisons. Measurement invariance will be assumed for the observed indicators conditional on the latent variables across countries. Various model extensions are proposed here to model the missing data mechanism together with the measurement and structural model. The model for the missing data mechanism will serve two purposes: first to characterize the item non-response as ignorable or non-ignorable and consequently to study the patterns of missingness and characteristics of non-respondents across countries but also to study the effect that a misspecified model for the missing data mechanism might have on the structural part of the model.

Exploring Reaction Time Distributions Through Hierarchical Bayesian Models

Poster Presentation

William Muntean, *Pearson*

Joe Betts, *Pearson*

Test fatigue is a powerful nuisance variable in cognitive experimental designs. Elongating retrieval tasks through repeated measures causes systematic variability in observations collected after a long unimpeded episode. This leaves experimenters with two alternatives. They can either reevaluate the methodological design or control effects via statistical models. The same measurement principles apply to high-stakes assessments: Over time, test fatigue contaminates measurement of latent traits. When reevaluation of the test design is not possible, statistical analyses of collateral information (e.g., reaction times across testing serial position) are a useful method to investigate the nuisance of test fatigue. The hierarchical framework of the first approach, inspired by van der Linden (2007), imposes a multivariate structure on two sublevel measurement models. This method conceptualizes test fatigue as systematic reduction of expected reaction times across variables of interest. Using a different hierarchical structure, the second approach, inspired by Rouder, Lu, Speckman, Sun, and Jiang (2005), models a test-taker's ability and an item's difficulty through a linear combination and incorporates a shift parameter in modeling item latency distributions. This method conceptualizes test fatigue as systematic changes in the reaction time distributions across the variables of interest. The utility of these approaches will be described.

Advances in Mixture Modeling

Keynote Address

Bengt Muthen, *Mplus*

After a brief overview of the many uses of finite mixture modeling, applications of new mixture modeling developments are discussed. One major development goes beyond the conventional mixture of normal distributions to allow mixtures with flexible non-normal distributions. This has interesting applications to cluster analysis, factor analysis, SEM, and growth modeling.

The talk focuses on applications of Growth Mixture Modeling for continuous outcomes that are skewed. Examples are drawn from national longitudinal surveys of BMI as well as twin studies. Extensions of this modeling to the joint study of survival and non-ignorable dropout are also discussed.

The Impacts of Ignoring a Crossed-Classified Covariate on Propensity Score Estimation

Paper Presentation

Ling Ning, *Texas A&M University*

Wen Luo, *Texas A&M University*

Felix J. Thoemmes*, Cornell University*

In this study, we consider propensity score estimation under unconfoundedness when data have a Cross-classified multilevel structure where the treatment is administered at the individual level and contextual (cross-classified cluster-level) variables influence the assignment mechanism and the outcomes. We simulated two-level cross-classified data structure to investigate the effect of ignoring a crossed covariate on propensity score estimation. We fit the simulated data with four approaches using logistic analysis of covariance for propensity score estimation: a) cross-classified random effects models, b) hierarchical linear random effects models that treat the data as strictly hierarchical, c) logit models with dummy variables for one crossed cluster, regarding the cluster effects as fixed parameters, but ignoring the other crossed cluster, d) logit models that ignore the cross-classified clustering as a whole. Preliminary results showed that when the degree of cross-classification is very low, the relative bias and RMSE of the Average Treatment effect on the Treated (ATT) were acceptable for the propensity score matching procedures for all except logit models that ignore the cross-classified clustering as a whole; however, when the degree of cross-classification is high, the performance of hierarchical linear random effects models and logit models with dummy variables are not acceptable.

Archetypal Analysis with Dimension Reduction

Paper Presentation

Yutaka Nishida, *Osaka University*

Archetypal analysis is a method of multivariate analysis which approximates a data matrix using few representative points called archetype. Archetype means extreme data points in data range.

Archetypal analysis has the same formulation as K-means clustering. In K-means clustering algorithm cluster representative points become its centroid. On the other hand, the representative points of archetypal analysis are more emphasized feature points. In clustering area, there is a procedure called tandem analysis. Researchers may conduct cluster analysis to the score of first few components, after carrying out principal component analysis to the data matrix for dimension reduction. This two-step procedure has a problem for clustering, because the subspace defined by the first few components does not reflect the true cluster structure. Similarly, it is possible to perform archetype analysis to the score of first few components. However, this method does not work for the same reason as the tandem analysis by cluster analysis. In this study we present a new method of performing dimension reduction and extraction of archetype simultaneously. The usefulness of this proposal method is shown using data.

Evaluation for Korean Automatic Scoring System (KASS) for Short Answers

Poster Presentation

Eun-Hee Noh, *Korea Institute for Curriculum and Evaluation*
Kyung-Hee Sung, *Korea Institute for Curriculum and Evaluation*
EunYoung Lim, *Korea Institute for Curriculum and Evaluation*

The purpose of this research is to evaluate an automated scoring engine for short answers in Korea, which has been developed using natural language processing techniques. Korean is different from English in terms of the characteristics and grammatical structure of language; English is an inflectional language and Korean is an agglutinative language. Moreover short answers focus on the content more than grammar and writing styles that are important criteria to score essays. Thus, Korean automated scoring engine has been developed for constructive responses, especially for short answers that are composed of 1~10 words. About 3,010 students' responses to the 38 short answer items in National Assessment of Educational Achievement (NAEA) have been used to evaluate KASS based on three criteria; the correlations, Kappa, and consistent rates between KASS and human raters' scores. For the results, the two criteria are higher than 0.64~1.00 and KASS works well for short answers. Also KASS can enhance accuracy of scoring, since KASS detected the scoring mistakes of human raters. However, as the short answers are getting more complicated, the correlations of scores between KASS and human raters decreased and decision making process by human was required more to improve the accuracy of scoring.

An Application to a Class Evaluation Questionnaire of Capture Rate

Paper Presentation

Kotaro Ohashi, *Rikkyo University*
Hideki Toyoda, *Waseda University*

The calculation method of the Capture Rate, an index standing for how many kinds of ideas were collected about class evaluation data which were written in a free description form was suggested. The DeLury method which was a popular calculation method of estimation of biological population was used for a calculation and a parameter of the number of the kinds of ideas was estimated. In addition, by using the phantom variables, it was shown that this method can be expressed as a structural equation model. The free description data of class evaluation coded in KJ method of a university was used for analysis. The capture rate was calculated and it became clear that the kinds of ideas more than the 80 % were collected as a result.

The Hybrid Multidimensional Item Response Model

Paper Presentation

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

There exist two major types of multidimensional item response theory (MIRT) models: compensatory and noncompensatory. The difference between them lies in their item response functions (IRF). Previously, researchers had to choose which model to use before applying it to data. In this study, a hybrid item response function, which is expressed as the weighted sum of a compensatory and a noncompensatory IRF, is introduced. Also, the estimation methods for the item parameters, namely, a set of compensatory and noncompensatory item parameters and a weight for each item, are developed. The existing compensatory and noncompensatory IRFs are now seen as two extremes of a continuum, with many variations in between. Therefore, using the proposed approach, researchers no longer have to subjectively choose either compensatory or noncompensatory IRF to apply. Instead, the proportion of each component is now estimated from the information within the data. Furthermore, since the hybrid IRF has a noncompensatory IRF embedded in it, the rotational indeterminacy of the multidimensional theta scale can be eliminated. The simulation study showed that with a suitably large sample size, the proposed method can recover the true structure well.

Assessing Cognitive Processes in Mathematics: Use of Sequential Items Based on Computer-Based Testing

Poster Presentation

Tomoya Okubo, *The National Center for University Entrance Examinations*

In this presentation, we demonstrate the utility of some types of sequential items based on computer-based testing. Although multiple-choice are considered to pose a difficulty in assessing cognitive processes, such items are often utilized in large scale testing because they offer ease in responding. On the other hand, though essays are subject to variability due to extraneous serendipitous factors, they are sometimes used to assess cognitive processes. Further, essay items incur a higher cost for recoding the responses, as compared multiple-choice items, which makes it difficult to use them in large-scale testing. Variations of items have been proposed to overcome these problems, such as the Modified Essay Question (MEQ); however, the utility of some item types can be improved by using computers. Thus, we propose the use of new sequential items based on computer-based testing, specifically for testing in mathematics, which enables us to assess a part of the cognitive process involved in completing these tasks. We also present the validation results of these proposed items.

A Comparison of Differential Item Functioning (DIF) Detection for Dichotomously Scored Item by Using IRTPRO, BILOG-MG, and IRTLRDIF

Paper Presentation

Mei Ling Ong, *University of Georgia*
Seock-Ho Kim, *University of Georgia*
Allan Cohen, *University of Georgia*
Stephen Cramer, *University of Georgia*

This paper is related to statistical issues of differential item functioning (DIF). Some previous social science studies considered all minorities as a homogeneous group. For instance, several studies mentioned that racial differences in assessment have primarily been developed in reference to comparisons between Whites and minority groups that include Blacks, Asians, Hispanics, and Native Americans. However, there is no evidence that Blacks and Hispanics are similar. Thus, DIF detection in this study is separately performed for ethnicity, White vs. Black, White vs. Hispanic, and White vs. Multi-Racial. The purpose is to determine which items contain bias for a specific race/ethnicity. This study has several goals. First, it presents an empirical data comparison of three programs, IRTPRO 2.1, BILOG-MG 3, and IRTLRDIF v.2, to detect DIF across majority and minority groups with the one-parameter logistic (1PL), the two-parameter logistic (2PL), and three-parameter logistic (3PL) models. Second, it examines IRTPRO to determine effectiveness in detecting DIF. Third, this study considers whether DIF exists for different ethnicities in the Georgia High School Graduation Predictor Test (GHSGPT). The initial results suggest that DIF exists in the Georgia High School Graduation Predictor Test by using the three programs.

On the Need of Reporting Precision of Reliability Estimates

Paper Presentation

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

Standard errors for test-score reliability are available but test constructors rarely report them in addition to the point estimates of reliability. This study estimates and adds confidence intervals to reliability estimates reported in the literature on test construction, and illuminates whether reliability reported for real tests can be trusted or depends so heavily on small sample size that conclusions about reliability should be reconsidered based on additional standard error information. We used the data base provided by the Dutch committee (COTAN; Evers, Sijtsma, Meijer, & Lucassen, 2010) in charge of test assessment in the Netherlands and Flanders. We retrieved data on 517 psychological tests available from 1997 onwards. Using sampling theory of reliability statistics, confidence intervals for coefficient alpha and other estimates were obtained. Rules of thumbs such as reliability must exceed .7 or .8 were reassessed. We found that regularly reliability estimates are insufficiently precise to justify claims that reliability exceeds the rules of thumb, when reliability is assessed for subpopulations.

Initial Condition Specifications in Autoregressive Latent Trajectory Models

Paper Presentation

Lu Ou, *The Pennsylvania State University*
Linying Ji, *The Pennsylvania State University*
Sy-Miin Chow, *The Pennsylvania State University*

Bollen and Curran (2004; see also Curran & Bollen, 2001) presented the Autoregressive Latent Trajectory (ALT) model as a synthesis of the autoregressive (simplex) model and the latent trajectory (curve) model. Due to the recursion in the ALT model, it has to be started up in practice. A challenge when estimating ALT models is to decide how to specify the initial conditions associated with the first available time point. Among the initial condition specifications (ICSs) considered in the literature include the predetermined and ongoing endogenous specifications (Curran & Bollen, 2001; Bollen & Curran, 2004; Hamaker, 2005; Jongerling & Hamaker, 2011). The former treats the first time point as predetermined and allows the process under study to conform to different longitudinal structures prior to, and after the first collected observation; the latter is used to describe an ALT process that has started in the distant past. One other specification that is commonly assumed in empirical applications is that the ALT process simply starts at the first time point and has no prior history. Unfortunately, the rationales for adopting one of these specifications over the others, their conditions of equivalence, and the effects of using covariates to predict individual differences in the presence of different ICs are poorly understood. In this study, we show analytically that a variety of ICSs is in fact nested within the predetermined ICS. We further demonstrate how likelihood ratio tests may be used to test hypotheses concerning ICSs with and without covariates using a subsample (N=3995) of longitudinal family income data from the National Longitudinal Survey of Youth (Bureau of Labor Statistics, 2012).

Item Analysis and Differential Item Functioning of A Coping Style Scale for Adolescents Using Bifactor Analysis

Paper Presentation

Xiang Zi Ouyang, *Beijing Normal University*
Tao Xin, *Beijing Normal University*
Fu Chen, *Beijing Normal University*

In this study we examine the psychometric characteristics of a Coping Style Scale for Adolescents questionnaire in Chinese teenagers using generalized full-information item bifactor analysis (see Li Cai, Ji Seung Yang, and Mark Hansen, 2011). A total of 1138 subjects of average age 16.88 years were recruited to complete the 34-item Coping Style Scale for Adolescents. Scale internal structure, item difficulty and discrimination, test information, and differential item functioning (DIF) were examined. The generalized bifactor model demonstrated good fit for the theoretical internal structure of the Coping Style Scale for Adolescents. The item difficulty is suitable for the subjects, and the majority of items discriminated at a moderate to high level along both general and domain-specific latent dimensions, and DIF by gender and grade was detected. Bifactor analysis provided a good fit for the Coping Style Scale for Adolescents and provides a comprehensive method of psychometric evaluation of the scale. The dimensionality of the scale is consistent with clinical theory. These findings provide support for the use of multidimensional modeling to examine item characteristics of Coping Style Scale for Adolescents.

The Investigation of the b Parameter Distribution and Sample Size on the Performance of Characteristic Methods

Paper Presentation

Eren Halil Ozberk, *Hacettepe University*
Akif Avcu, *Marmara University*
Hulya Kelecioglu, *Hacettepe University*

The aim of the current study was to investigate the performance of two commonly used characteristic curve equating methods (Stocking-Lord and Haebara) with varying b parameter distributions and sample sizes. Population parameters were generated with varying b parameter distributions (skewness_b. = -0.5/0/0.5). The descriptive statistics of the population parameters for the Base Test and Equated Test were specified. Additionally, the size of the sample was set at 7 different levels (150, 250, 500, 750, 1000, 1500, 3000). By using those ability and theta vectors, response matrices was generated where 80 unique items and 20 common items were used. Finally, this response generation operation was repeated 100 times for each condition. The performance of the CCM's was tested by using RMSE. For the conditions where the b parameter is not distributed normally, RMSE values are not decreasing monotonously with increasing sample size. Additionally, for both methods, different b parameter distributions give different results for lower sample sizes but as the sample size increases, the results become similar. Finally, even though the differences were ignorable, the Haebara Method yielded less error for the b parameter for smaller sample sizes in all conditions. The results are discussed in the context of the relevant literature.

Examining Testlet Effects in English Proficiency Test: A Bayesian Testlet Response Theory Approach

Paper Presentation

Burhanettin Ozdemir, *Hacettepe University*

A testlet is defined as a group of items based on the same stimulus (Wainer and Kiely, 1987) and used commonly in language testing. However, testlets generally violate local independence which is an important assumption under the item response theory (IRT) model. One direct way to handle the local dependence effect is by adding a testlet effect term to the IRT model which is called testlet response theory (TRT) model. In this study, the listening section items and gap-filling test items embedded in the reading passages of English Proficiency Test (EPT) were analyzed to determine the degree of testlet effect and its outcomes with two different Bayesian testlet response theory approaches using MCMC methods. In addition, results of IRT model which consider each item in testlets to be local independent were compared to results of two different TRT models with respect to estimates of items and ability parameters. The testlet data set from EPT which was administered by Hacettepe University in June 2011 was used. Totally, 1126 university students participated in 2011 EPT. Two listening passages (i.e., two testlets) with 7 items and seven gap-filling passage testlets were analyzed using a three-parameter logistic testlet response model and graded response model. Results indicated small testlet effects for listening testlets, while moderate to large effects for the gap-filling testlets. Standard IRT analysis led to an underestimation of ability estimates and an overestimation of the standard error of ability estimates when testlet effect was neglected. Although item parameter estimates remained largely unaffected, standard IRT analysis underestimated item discrimination and pseudo-chance parameters while overestimated item difficulty compared to TRT model. Both Implications for the analysis and evaluation of testlet-based tests were discussed.

Longitudinal Group-Level Diagnostic Assessment for Korean Learners of English

Paper Presentation

Chanho Park, *Keimyung University*

Sookyung Cho, *Hankuk University of Foreign Studies*

Obtaining fine-tuned diagnostic information of individuals by applying cognitive diagnosis models can be useful when tailoring instruction for students' needs; however, diagnosis of groups can be necessary when groups are compared for various reasons. For example, educational policies are often made for specific groups, and group-level diagnosis can provide information for the effectiveness of the policies. Also, such assessments as the Trends of International Mathematics and Science Study are administered to compare countries. Attempts have been made to develop models for group-level comparisons (Park & Bolt, 2008; Prowker & Camilli, 2007; Tatsuoka, Corter, & Tatsuoka, 2004), and group-level diagnostic assessment can provide richer information when such information is accumulated across years. In this study, a group-level diagnostic assessment model based on multilevel item response theory is illustrated and applied across three years to the English test data of the National Assessment of Educational Achievement (NAEA), which is a national assessment for Korean students' educational achievement. The results show how the profiles of provinces and metropolitan cities of Korea change across three years for the knowledge, skills, and abilities necessary for the NAEA English test. Methodological issues of such approaches are also discussed.

Developments of Automated Essay and Short-answer Scorers

Poster Presentation

Doyoung Park, *Korea Institute for Curriculum and Evaluation*
Kija Si, *Korea Institute for Curriculum and Evaluation*
Hwanggyu Lim, *Korea Institute for Curriculum and Evaluation*
EunYoung Lim, *Korea Institute for Curriculum and Evaluation*

KICE Essay Scorer (KES) and KICE Short-answer Scorer (KSS) were developed and enhanced by the Korea Institute for Curriculum and Evaluation in order to improve the scoring reliability and efficiency of the National English Ability Test writing section. KES adopted machine learning algorithms such as maximum entropy, support vector classification, and support vector regression, whereas KSS included support vector regression, Bayesian network, and multilayer perceptron algorithms. And extensive analyses and calibration of scoring features and models resulted in 152 and 37 features for KES and KSS, respectively. To evaluate the performances of KES and KSS, four sets of real data scored by standard score experts, human raters, and these two machine scorers were rigorously analyzed and compared in terms of item characteristics, inter-rater reliability, test score reliability measured by Generalizability theory, and rater severity measured by the Multi-facet Rasch model. Both KES and KSS showed no significant differences in item difficulties and discriminations, inter-rater correlations, exact and adjacent agreements, and test score reliability when compared with human raters. Differences in rater severity were found between human and machine scoring, however. KES and KSS tended to have rater severities whose effects were far less minimal than human scoring.

Penalized Likelihood Principal Component Analysis

Paper Presentation

Trevor Park, *University of Illinois at Urbana-Champaign*

Exploratory principal component analysis (PCA) suffers from two problems: components can be difficult to interpret, and they are subject to high sampling variability. Penalized likelihood PCA, like many recently-proposed regularization methods, employs a penalty function to enhance interpretation and stability. Unlike other methods, it uses a standard statistical framework, which keeps the biased components more faithful to the data and provides standard tools for inference. We examine how penalized likelihood PCA treats the problem of high sampling variability of the components. We briefly discuss how the penalty can be tailored to suit particular applications. We also briefly consider efficient computation using a modern algorithm for orthonormally-constrained optimization.

The Science of Cause and Effects

Keynote Address

Judea Pearl, *UCLA*

The enormous progress in the science of causation in the past two decades has hardly impacted psychometric practices and education. To close this gap, I will reduce the science to just two fundamental principles:

1) How counterfactuals are to be computed from an assumed data-generating process, and

2) How dependencies in the observed data can be inferred from the structure of the process.

I will then describe how these two principles alone lead to remarkably powerful techniques of answering practical, yet non-trivial questions in psychometric research. These include: policy evaluation, confounding control, external validity, sample selection bias, heterogeneity, mediation, causes-of-effects and missing data.

Detecting Nonlinearity of Latent Relationships with Confidence Envelopes for a Semiparametric Approach to Modeling Bivariate Nonlinear Relations among Latent Variables

Paper Presentation

Jolynn Pek, *York University*

R. Philip Chalmers, *York University*

The functional form of a latent regression is typically unknown during early phases of research, highlighting the advantage of using semiparametric models (SPM) over parametric counterparts to model potential nonlinearity. An indirect application of structural equation mixture models (Bauer, 2005) can flexibly recover and describe the form of the unknown latent relationship with minimal distributional assumptions. To evaluate potential nonlinearity of the latent function as a whole, approximate simultaneous confidence bands or confidence envelopes (CEs), based on the delta method and parametric bootstrap, are developed and evaluated by Monte Carlo for coverage. Additionally, a line finding algorithm to be used in conjunction with these CEs is developed as an implementation of an informal test to diagnose nonlinearity. Targeted simulations were used to evaluate performance of the algorithm in terms of rates of detecting nonlinearity. Recommendations for the use of these CEs and the algorithm for detecting nonlinearity are suggested.

Score-Tests for Detecting Differential Item Functioning in Cognitive Diagnosis Models

Paper Presentation

Michel Philipp, *University of Zurich*
Achim Zeileis, *University of Innsbruck*
Carolin Strobl, *University of Zurich*

Differential item functioning (DIF) is a relatively new topic for cognitive diagnosis models (CDM). Recent research in this area has suggested various approaches for DIF detection, that are designed for comparison between given groups (such as males and females). In this talk, we will present work in progress on a new DIF detection approach for the DINA model that is based on score-tests. These tests are able to identify structural changes in the individual score contributions. They can be used to detect parameter differences between given groups suspected of DIF, such as males and females, but are also computationally feasible for detecting parameter differences along continuous variables, such as age, without previous discretization. We will illustrate the properties of the new DIF detection approach by means of simulation results and discuss its potential for future integration into model-based trees. Moreover, the method is not restricted to the DINA model and we plan an extension to the generalized DINA model framework in future research.

Simulating the Impact of Acquiescence Response Style

Poster Presentation

Hansjörg Plieninger, *University of Mannheim*

Response styles are defined as the tendency to respond to questionnaire items irrespective of content. There is a widespread claim and fear that response styles, such as extreme responding or acquiescence, threaten the quality of self-report data. The goal of the present study was to take a more systematic view of this issue. Therefore, a simulation study was carried out to scrutinize the effect of ignoring response styles in applied data analyses. Data were generated with a multidimensional Rasch model using different weights for the content- and response style-related dimensions, respectively. The analyses focused, for example, on the effect of acquiescence on the correlation of two scale scores and investigated the effect of factors such as amount of response style variance. The results show that acquiescence indeed can distort results. However, this negative effect can be controlled for to a large extent by the use of reverse-coded items. In summary, the results may serve a more systematic view on the role of response styles and may guide the design of self-report questionnaires.

Item Latent Class Response Model for Guessing Behaviors in Data with Timing Information

Paper Presentation

Artur Pokropek, *Educational Research Institute*

In a situation of reduced motivation proper estimation of latent traits and items parameters might be difficult. The probability of a correct response in such settings would be dependent not only on real traits of respondents but also on individual motivation. When timing information is available some corrections for estimation in such settings are possible, for instance item filtering according to response time. In this paper a new solution is proposed, an Item Latent Class Response Model (ILCRM) that would incorporate information about timing during the estimation of model parameters. We refer to this model as "item latent class" because latent classes are defined on the item level rather than on the individual level like in most of the mixture IRT models. In this work the model is formally introduced. A simulation study, reflecting different low stakes conditions, is shown proving that in certain conditions ILCRM reduced biases in estimates of model parameters. Finally some real data examples are presented.

Comparing the Effectiveness of Rater Training Methods Using Generalized Linear Mixed Models

Paper Presentation

Kevin Raczynski, *The University of Georgia*
Allan Cohen, *The University of Georgia*
George Engelhard, *The University of Georgia*
Laura Lu, *The University of Georgia*

There exists a large body of research on the effectiveness of rater training programs in the Industrial and Organizational Psychology literature. Far less research has been done in the context of educational assessments featuring constructed responses. The purpose of this study is to compare the effectiveness of two widely used rater training methods—self-paced and collaborative frame-of-reference—in the context of a large-scale, statewide writing assessment. In this study, 66 raters were randomly assigned to the training methods. After training, all raters scored a common set of fifty representative essays. To determine raters' accuracy on these essays, raters' scores were compared to resolved expert scores and coded accurate (1) when the scores matched and inaccurate (0) otherwise. This approach was taken because over ninety-nine percent of these comparisons aligned either exactly or within one point. A series of generalized linear mixed models were then fitted to these binary data. Results suggested that the self-paced method was equivalent in effectiveness to the more time-intensive and costly collaborative method. Implications for large-scale educational assessments and suggestions for further research are discussed.

On Optimal Shortening of Psychometric Scales

Paper Presentation

Tenko Raykov, *Michigan State University*

A latent variable modeling procedure for optimal shortening of multi-component measuring instruments is outlined, which is based on latent construct predictability by their observed components. The approach can be used when an original scale, or section of it, needs to be shortened without considerable loss in predictive power with respect to the underlying construct. The method evaluates the critical parameters on which the contributions of individual components to the instrument's maximal reliability coefficient depend, as well as the loss in that coefficient associated with instrument shortening. The procedure is readily employed in empirical research using popular software, and is illustrated with a numerical example.

Study the Quality of Items using Isotone Nonparametric Regression in a Mathematics Test

Paper Presentation

Pilar Rodriguez, *Universidad de la República*
Mario Luzardo, *Universidad de la República*

We discuss the use of isotone nonparametric item response theory models based on nonparametric regression to analyze the quality of items in a mathematics test to first generation students of higher education. In the context of our approach we take a grid $0, \frac{1}{T}, \dots, \frac{i}{T}, \dots, 1,$ and use the nonparametric estimate of the ICC in each point

$$\widehat{P^{-1}}(\theta) = \frac{1}{Th_d}\int_{-\infty}^{\theta}\sum_{i=1}^{T} K_d\left(\frac{\widehat{P*}\left(\frac{i}{T}\right) - u}{h_d}\right)du$$

where K_r and h_r are the kernel and the bandwidth for the regression estimator respectively. The inverse of the monotonous ICC in θ will be

$$\widehat{P*}\left(\frac{i}{T}\right) = \frac{\sum_{j=1}^{N} K_r\left(\frac{\frac{i}{T} - \widehat{\theta_J}}{h_r}\right) Y_j}{\sum_{j=1}^{N} K_r\left(\frac{\frac{i}{T} - \widehat{\theta_J}}{h_r}\right)}$$

where K_d and h_d are the kernel and the bandwidth for the isotone estimator. The estimate of $\hat{P}$ is obtained by reflexion of $\widehat{P^{-1}}$ with respect to the line $y = x$. We compare these models with parametric IRT models.

Can the Homoscedasticity Assumption Be Ignored? To Assume or Not to Assume

Poster Presentation

Patrick J. Rosopa, *Clemson University*
Alice M. Brawley, *Clemson University*
Theresa P. Atkinson, *Clemson University*
Stephen A. Robertson, *Clemson University*

For optimal performance, many statistical tests require assumptions to be satisfied, including homoscedasticity (Fox, 2008). Some research, however, suggests that preliminary tests for homoscedasticity may be unnecessary (Sawilowsky, 2002; Zimmerman, 2004). The present Monte Carlo study compares two diagnostic tests for heteroscedasticity (Breusch & Pagan, 1979; Levene, 1961) and several statistics for testing mean and slope differences. We compare these statistics when used without a preliminary test for homoscedasticity (i.e., unconditional Type I error and power) and with a preliminary test for homoscedasticity (i.e., conditional Type I error and power, depending on results of the diagnostic tests). For mean differences, when heteroscedasticity was directly paired, conventional tests performed poorly; Welch's test (used both unconditionally and conditionally) and unconditional weighted least squares (WLS) regression controlled Type I error and had adequate power. When heteroscedasticity was indirectly paired, unconditional Student's t outperformed the other tests in terms of

Type I error and power to detect mean differences. When testing slope differences, OLS regression showed the best power, but inflated Type I error rates. Conditional use of WLS regression was balanced in terms of control of Type I error rate and adequate power for slope differences. Recommendations for researchers and practitioners are discussed.

Exploratory Latent Curve Modeling with Fractional Polynomials

Paper Presentation

Ji Hoon Ryoo, *University of Virginia*
Dingjing Shi, *University of Virginia*

In the cross-sectional design, there are four different models including factor analysis models, item response models, latent profile models, and latent class models in terms of types of variables (Collins and Lanza, 2010). The corresponding models for longitudinal data can be considered as latent curve models for continuous observed variables, latent curve models for categorical observed variables, mixture growth models, and latent transition analyses, respectively. To include such measurement errors in the regression-type of analyses, Laird and Ware (1984) introduced a linear mixed effects model (LMM) that can also be viewed as a latent curve model (LCM) in the SEM framework. The LCM is more flexible than the LMM when we consider latent variables measured by observed variables for each construct (Hsieh, von Eye, Maier, Hsieh, & Chen, 2013). However, there is still room to develop the LCM using flexible functional forms in LMM. For example, the functional form for growth patterns can also be exchanged between the LMM and the LCM literature (Grimm, Steele, Ram, & Nesselroade, 2013) when we consider the exploratory analysis (Asparouhov & Muthen, 2009). In this study, we introduce a non-linear functional form, fractional polynomial (Long & Ryoo, 2010), to the LCM.

The Impact of Model Misspecification with Multidimensional Test Data

Paper Presentation

Sakine Gocer Sahin, *Hacettepe University*
Cindy M. Walker, *University of Wisconsin – Milwaukee*
Selahattin Gelbal, *Hacettepe University*

In this study, we simulated data for 5000 examinees on a thirty item two-dimensional test, using a compensatory MIRT model. Various combinations of simple and complex structured items were examined. Specifically, the numbers of simple structured items on the tests were gradually decreased from 24 to 6, in multiples of six, while simultaneously increasing the number of complex items by the same number of items. In one scenario, the simple structured items were simulated to measured both dimensions equally. In a second scenario, the simple structured items were simulated to measure only the first dimension. Other variables in the study included the correlation between dimensions, and varying the ability distributions on the first and second dimensions. RMSE and correlations were used to determine the impact of model misspecification, using the results of a unidimensional simulated and scaled test for comparison purposes. Preliminary results indicate that the underlying structure of multidimensional tests does have an impact on estimation error. Specifically, in some instances fitting a unidimensional model to multidimensional data results in estimation error that is not much different from what is obtained when fitting a unidimensional model to unidimensional data.

Adjusting for Many Covariates in a Matching-Based Educational Program Evaluation

Paper Presentation

Adam Sales, *Carnegie Mellon University*
Ben B. Hansen, *University of Michigan*

This paper will suggest an approach to evaluating educational interventions that supplements covariate matching with modern high-dimensional prediction or machine-learning. The method attempts to protect matching estimates against the bias that would result from omitting a measured covariate from a matching scheme. After constructing a match using a limited number of covariates, selected for their theoretical importance, a researcher would then use a high-dimensional outcome regression on the entire set of covariates to estimate prognostic scores' predictions of the outcome of interest as a function of the full covariate matrix. These can be used to test balance, as well as to further adjust the causal estimate and reduce confounding bias. In some circumstances, they will decrease the variance of the estimate as well. We will present theoretical results to justify the method's bias-reducing properties, as well as a simulation study. Additionally, we will demonstrate the method in an evaluation of a school-level intervention that took place in six Kentucky high schools.

Latent Class Analysis of Multidimensional Alternative Constructs

Poster Presentation

Kelli Samonte,*University of North Carolina at Greensboro*
Terry Ackerman, *University of North Carolina at Greensboro*

Often, assessments aim to measure respondents' ability level across a range of proficiency levels (i.e., basic, proficient, advanced). Mislevy (1992, p.15) notes that, "contemporary conceptions of learning do not describe developing competence in terms of increasing trait values, but in terms of alternative constructs." That is, instead of simply gathering more and more knowledge on one, general, construct, it may be that constructs evolve as knowledge increases. The current proposal aims to examine the extent to which latent class analysis is able to correctly classify individuals to their respective proficiency level given the dimensionality of their scores. This is confirmatory in the sense that the dimensionality of each proficiency level is treated as known. The conditions for the simulation will vary the number of people per group (equivalent groups vs. non-equivalent groups) and the correlations between the factors (low, medium, and high). The number of people per group condition was chosen to examine how well the model performs when the group sizes vary (e.g., a small number of "advanced" individuals). The correlation among factors condition was chosen to see how well the model performs as the factor structure more closely resembles a unidimensional scale.

The Infinitesimal Jackknife and Analysis of Higher Order Moments

Paper Presentation

Albert Satorra, *Universitat Pompeu Fabra, Barcelona*
Robert Jennrich, *University of California, Los Angeles*

Higher order moments are expressed as expectations of Kronecker products of centered random vectors. It is shown that sample versions of these higher order moments are asymptotically normal and shown how to use the infinitesimal jackknife to estimate the asymptotic covariance matrices of these sample versions. It is shown how to use these to test the goodness of fit of a higher order moment structure model and estimate standard errors for its parameters. Applications include nonlinear factor analysis and errors in variables models.

Regularization Methods and the Modeling of Differential Item Functioning

Paper Presentation

Gunther Schaubergerm, *Ludwig-Maximilians-Universität München*
Gerhard Tutz, *Ludwig-Maximilians-Universität München*

A new diagnostic tool for the identification of differential item functioning (DIF) is proposed. Classical approaches to DIF allow for considering only a few subpopulations like gender or ethnic groups when investigating if the solution of items depends on the membership to a subpopulation. As proposed in Tutz and Schauberger (2013), we set up the following model

$$\log\left(\frac{P(X_{pi}=1)}{P(X_{pi}=0)}\right)=\theta_p - \left(\text{ß}_i + x_p^T Y_i\right)$$

when person p tries to solve item i. It includes the usual person ability θ_p but the difficulty of the item is given by a linear term that contains a vector of person-specific variables x_p, which can contain metric as well as categorical components. The vector of variables contains the potential candidates for inducing DIF. A consequence of the flexibility of the model is that it contains a large number of parameters. It is shown that penalized maximum likelihood estimators and boosting techniques can be used to solve the estimation problem and to identify the items that induce DIF. It is demonstrated that the method is able to detect items with DIF. Simulations show that the method competes well with traditional methods of DIF detection. The proposed method is applied to data from an intelligence test.

Predictive Inference Using Latent Variables with Covariates

Paper Presentation

Lynne Steuerle Schofield, *Swarthmore College*
Brian Junker, *Carnegie Mellon University*
Lowell Taylor, *Carnegie Mellon University*
Dan Black, *University of Chicago*

Plausible Values (PVs) are a standard multiple imputation tool for analysis of large education survey data that measures latent proficiency variables. When latent proficiency is the dependent variable, we reconsider the standard institutionally-generated PV methodology and find it applies with greater generality than shown previously. When latent proficiency is an independent variable, we show that the standard institutional PV methodology produces biased inference because the institutional conditioning model places restrictions on the form of the secondary analysts' model. We offer an alternative approach that avoids these biases based on the mixed effects structural equations (MESE) model of Schofield (2008).

Studying the Interplay Between Familial Effects and School Influence in PISA dData

Paper Presentation

Inga Schwabe, *University of Twente*
Stephanie M. van den Berg, *University of Twente*
Martina R.M. Meelissen, *University of Twente*
Annemiek R.A. Punter, *University of Twente*

Using twin data, observed variance in family members can be decomposed into genetic variance, shared-environmental variance and non-shared variance. The familial effects on, say, a test score may reflect genetic effects, shared environmental effects, or a combination of both. A considerable part in educational test scores is due to familial effects (genetic and shared-environmental effects). For IQ it has been shown that the non-shared environmental variance is dependent on the familial effect: the influence of non-shared environmental effects is smaller for children from high-scoring families, suggesting that a familial propensity for high IQ makes a child relatively immune to random environmental circumstances. It would be interesting to test for a similar pattern in educational test scores. Here we propose a method to test for such an interplay of environment and talent. In the absence of twin data, we argue that socio-economic status (SES) of a family can serve as proxy for the familial effect on school achievement and that this effect may predict the impact of the school effect. The method is applied to PISA data. We hypothesize that the school influence is less for children from low and high SES families than for children from average SES families.

On the Reasons of Detecting a Difficulty Factor in Confirmatory Factor Analysis of Binary Data with Constrained Models: A Simulation Study

Paper Presentation

Karl Schweizer, *Goethe University Frankfurt*

So-called difficulty factors have repeatedly been observed in investigations of the structure of binary data. Such factors are considered as a reflection of the pattern of item difficulties. In order to learn about the possible influence of different item difficulties on the outcome of confirmatory factor analysis of binary data if the model of measurement is constrained, simulated data showing specific patterns of item difficulty were investigated. The data were constructed in assuming one underlying source of homogeneity and different ranges of item difficulty (broad, medium, small). The structure of the data was investigated by comparing constrained models with one and two latent variables. In assuming that the difficulty factor was nothing but an artifact resulting from random error it was expected that only in up to five percent of the outcomes the investigations would be in favor of a model with two latent variables. Unfortunately, the outcomes disproved the expectation: The chi-square difference led in all conditions to the detection of more latent variables than expected by chance. The overestimation was especially severe if the range of item difficulty was broad. The CFI difference yielded better results than the chi-square difference in ten or more manifest variables.

Model Selection for Multilevel Mixture Rasch Models

Paper Presentation

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

Item response theory (IRT) models assume that a single model applies to all examinees in the population. However, examinees may have response patterns which differ substantially enough to suggest that they come from different subpopulations. Although mixture IRT (MixIRT) models often can be used to handle the heterogeneity among the individuals in different subpopulations, these models do not account for the multilevel structure that is common in much educational and psychological data. Multilevel extensions of MixIRT models have been proposed to address this shortcoming. Successful applications of both MixIRT models and their multilevel extensions depend on the detection of the best fitting model. Previous studies have compared performances of different model selection indices with MixIRT models. No such study, however, has been conducted for multilevel MixIRT models. In this study, performance of several information indices (i.e., AIC, BIC, CAIC, and SABIC) will be compared for model selection with a multilevel mixture Rasch model in the context of a simulation study. Two sample sizes (1,500 and 3,000), two test lengths (10 and 20), and three different class solutions for a multilevel mixture Rasch model will be manipulated in the simulation study. Fifty data sets will be simulated for each condition.

Test Speededness Detection based on the Detection of Change Point

Paper Presentation

Can Shao, *University of Notre Dame*
Jun Li, *University of Notre Dame*
Ying Cheng, *University of Notre Dame*

Not having enough time to fully consider the items will lead to smaller probabilities of answering questions correctly or larger proportion of missing responses at the end of a test. Test speededness often leads to biased estimates of item and ability parameters; successfully detecting and deleting speeded responses may help reduce this bias. In this study, we propose a change-point-detection-based method to detect if there is speededness and when an examinee starts to speed. The method uses an iterative procedure:(1) estimate item and ability parameters based on the full dataset; (2) for each examinee, identify the most likely position of the change point and calculate the likelihood-ratio statistic for speededness; (3) permute items to obtain the null distribution of the likelihood-ratio statistic, compare the observed statistic against the null distribution to estimate the false discovery rate(FDR); (4) use a pre-specified FDR cutoff to determine examinees suspected of speededness and remove their responses; (5) re-estimate item and ability parameters based on the "cleansed" dataset; (6) repeat steps (2)-(5) until the number of examinees detected as speeded does not change significantly. Simulation studies show that this procedure detects test speededness efficiently and dramatically reduces the bias of item and ability parameter estimation.

A GPU-based Gibbs Sampler for the Multi-unidimensional IRT Model

Poster Presentation

Yanyan Sheng, *Southern Illinois University Carbondale*
William S. Welling, *Southern Illinois University Carbondale*
Michelle M. Zhu, *Southern Illinois University Carbondale*

Fully Bayesian estimation shows promise for IRT models. Given that it is computationally expensive, the procedure is limited in actual applications. It is hence important to seek ways to reduce the execution time, and a suitable solution is to use high performance computing. Given the high data dependencies in a single Markov chain for IRT models, such as the dependency of one state of the chain to the previous state, and the dependencies among the data within the same state, the implementation of parallel computing is not straightforward. Previous studies suggest that when it comes to parallel computing, massive core-based graphic processing units (GPU) offer advantages and are more cost effective than the message-passing interface. This study focused on the development of the GPU-based algorithm to implement the Gibbs sampler for the multi-unidimensional IRT model, which is applicable for tests that involve multiple latent traits with each item measuring one of them. The results using one GPU card indicate that the developed parallel algorithm was efficient and could achieve a speedup of up to 20 times that of the serial implementation. This further sheds light on developing GPU-based Gibbs samplers for other multidimensional or more complicated IRT models.

Testing the Model Fit Index of Structural Equation Modeling: A Simulation Study to the Performance of the Unbiasedness of RMSEA

Paper Presentation

Dingjing Shi, *University of Virginia*
Ji Hoon Ryoo, *University of Virginia*

Model evaluation is essential in the development and application of SEM. To assess the model correctness, an abundance of alternative fit indices supplement the traditional χ^2 test (Hu & Bentler, 1999). Among them, the root-mean-squared-error-of-approximation (RMSEA) receives popularity and is believed to be an unbiased estimator (Hu & Bentler, 1998; Chen et al, 2008). Due to the data reduction nature, SEM model with high factor loadings was an assumed condition in RMSEA research (Curran et al, 2002). Empirical researchers unconditionally rely on RMSEA in reports even when the models have low factor loadings, thus drawing unreliable conclusions of the model fit. This paper intends to investigate the difference between methodological and empirical research by examining the unbiasedness property of RMSEA and its confidence intervals (CIs) when factor loadings are small (less than .5) in the model, which is not uncommon in practice. The study also answers the questions how RMSEA differs in performance as sample sizes change, and how it affects the power under such conditions. We present results from simulation studies using Mplus. Preliminary results show that RMSEA and its CIs are unbiased with low factor loadings, and the goodness-of-fit criteria might be conservative under some of these conditions.

The Introduction of the Strand-Level CAT Algorithm

Paper Presentation

David Shin, *Pearson*
Yuehmei Chien, *Pearson*

The purpose of this paper is to introduce the strand level CAT (SL-CAT) algorithm and compare it to the overall CAT algorithm. The features of the SL-CAT algorithm include: (1) Item selection can be based on either the strand-level theta or the overall theta or the combination of the two. Therefore, if students are more capable on certain strands and less capable on other strands, the items selected from those strands can possibly be based on their strand-level theta. (2) The item exposure (IE) rate can be controlled at the strand level. Therefore, if some strands in the item pool have more items than the other strands, the item exposure control can be set according to the strand pool size to more efficiently use the items in the pool. (3) The content balancing can be guaranteed at the strand level using the SL-CAT algorithm. In this paper, several operational item pools will be used to illustrate differences between the SL-CAT and the overall CAT algorithms. The item pools used for comparison will include pools with different IRT models, different pool sizes, and different correlations between strand-level thetas. Results will be evaluated by: ability estimation, IE rate, and content balancing.

Using Markov-IRT to Characterize Process Data

Paper Presentation

Zhan Shu, *ETS*
Alina von Davier, *ETS*
Mengxiao Zhu, *ETS*
Yoav Bergner, *ETS*
Jiangang Hao, *ETS*

In the Scenario-Based Tasks, students may be asked to make a series of decisions to solve a problem and/or achieve a goal. Therefore, the process data could be seen as a series of decisions made in situations where outcomes are partly random and partly under the control of a decision maker (Bellman,1957). More precisely, it is a discrete time stochastic control process. A Markov-based IRT has been built to capture certain features of the process data: [1] what actions/steps chosen by students, and [2] the transition from one action to another. In this paper, its model structure, parameter space and estimation will be discussed at first. Then, the application of the Markov-based IRT and the outputs of the model will be discussed under the framework of the evidence-central design (ECD) of the Wells, as an example of how the psychometric/statistical evidence could be used to support the cognitive hypothesis made in the ECD.

Goodness of Fit Methods for Nonparametric IRT Models

Paper Presentation

Klaas Sijtsma, *Tilburg University*
J. Hendrik Straat, *Cito Arnhem, the Netherlands*
L. Andries van der Ark, *University of Amsterdam, the Netherlands*

A general, flexible and much used IRT model is the monotone homogeneity model, of which most of the unidimensional, locally independent and monotone IRT models for dichotomous and polytomous items are special cases. Several methods exist to assess the goodness of fit of the monotone homogeneity model. We provide a brief overview of these methods, which will show that most methods focus on identifying subsets of unidimensional items or on assessing monotonicity of item response functions. Methods for assessing local independence seem to be rare. In addition to providing an overview, we discuss some recent research that uses the conditional association property (Holland & Rosenbaum, 1986) to assess local independence. Conditional association is implemented in a procedure that aims at identifying subsets of items, where the items within one subset are locally dependent but items from different subsets together are locally dependent. The new procedure was found to produce larger item sets than the alternative procedures DETECT and Mokken scale analysis. For use in real data analysis, we recommend a comprehensive goodness of fit package that assesses all assumptions of the monotone homogeneity model.

An Examination of Kernel Equating and the Test Characteristic Curve Method

Paper Presentation

Bradley Smith, *University of Nebraska – Lincoln*
R.J. De Ayala, *University of Nebraska – Lincoln*
Rebecca L. Norman Dvorak, *Human Resources Research Organization*

This study examined the accuracy of Kernel Equating (KE) and the Test Characteristic Curve method (TCC) through a Monte Carlo simulation using the non-equivalent anchor test (NEAT) equating design. Independent variables are sample size (levels: 1000, 2000, 10,000), test lengths (levels: 25, 50, 100), average form discrimination (levels: low, medium, high), different anchor reliabilities (levels: constant, variable), and the percent of anchor items conditions (levels: 10%, 20%, 30%). The dependent variable is the accuracy of simulees' equated scores on Form Y with respect to the simulees' parametric true scores (i.e., Root mean square difference (RMSD)). For each cell in the design 100 replications were conducted. The two parameter logistic (2PL) model was used for TCC equating and for calculating parametric true scores. For KE both chain equating (CE) and post-stratification equating (PSE) are examined. Preliminary results show that the accuracy of each equating method varied along the score range of Form X. Specifically, KE is most accurate for individuals who score in the middle to upper-middle range of scores on X. This corresponds to the range in which the majority of individuals fell. In contrast, TCC is more consistently accurate across the entire range of scores.

Applicability of Common Core State Standards in English Language Proficiency Assessment

Paper Presentation

Yoon Ah Song, *University of Iowa*
Jinah Choi, *University of Iowa*
Catherine Welch, *University of Iowa*

While the No Child Left Behind Act (NCLB, 2001) suggested the minimum standards for learning and teaching from the perspective of educational accountability, it can be viewed that Common Core State Standards (CCSS, 2009) was established for excellence in education in U.S. CCSS means that students have to achieve higher standards than the past in each of the subject area. It will improve the quality of education in U.S. However, it may be less likely that the implementation of the CCSS would bring the same improvement to English Language Learner (ELL) students; it could be more challenging and demanding. As every state is adapting the CCSS, there is little effort to reflect this CCSS to their English Language Proficiency (ELP) Standards and assessment. This study is focused on evaluating whether current state ELP standards and assessments are appropriate for ELL students to achieve CCSS like non-ELL students. To conduct the research, we compare ELP Standards of several states with the CCSS in both content level and item level. This will help educators predict performance change in ELL students and prepare appropriate steps.

Where is the log? A Comparison of the Prior Information Criterion and the Bayes Factor

Paper Presentation

Sara Steegen, *KU Leuven*

Priors, while often maligned for supposedly introducing subjectivity, can also be viewed more beneficially as opportunities to capture theory in a model. A logical consequence of this perspective is that model selection methods should be sensitive to the prior. Most existing model selection methods, however, pride themselves as being insensitive to the prior, the most notable exception being the Bayes Factor (BF). Another example is the recently proposed Prior Information Criterion (PIC), a modification of the Deviance Information Criterion for evaluating (in)equality constrained hypotheses (Van de Schoot, Hoijtink, Romeijn & Brugman, 2012). PIC is closely related to the BF, in the sense that the latter averages the likelihood as weighted by the prior, whereas PIC averages the log-likelihood as weighted by the prior. While in some situations PIC and BF behave similarly and result in identical conclusions, we show that in other situations PIC can lead to conclusions that not only widely differ from the conclusions based on the BF, but are also highly questionable.

Propensity Score Designs for Causal Inference: Challenges in Practice

State of the Art Talk

Peter Steiner, *University of Wisconsin- Madison*

The popularity of propensity score (PS) methods for estimating causal treatment effects from observational studies has been strongly increasing during the past decade. However, the success of PS designs in removing selection bias rests on strong assumptions for identifying and estimating causal effects—particularly the strong ignorability assumption that requires that all confounding covariates are reliably measured. After an introduction to the design and analysis of different PS methods (matching, stratification, weighting, and regression), the talk mainly focuses on practical challenges in implementing a valid PS design: (i) The selection of baseline covariates for removing confounding bias; (ii) the influence of covariate measurement error on bias reduction; (iii) the choice of a specific PS design; (iv) the assessment of hidden bias due to unobserved confounders. All these issues will be illustrated and discussed using from simulation studies, meta-analyses and within-study comparisons from the social and behavioral sciences.

Structural Equation Modeling Approaches for Analyzing Partially Nested Data

State of the Art Talk

Soyna Sterba, *Vanderbilt University*

Study designs involving clustering in some study arms, but not all study arms, are common in clinical treatment-outcome and educational settings. For instance, in a treatment arm, persons may be nested in therapy groups, whereas a control arm may have no groups. Methodological approaches for handling such partially nested designs have previously been developed in a multilevel modeling framework (MLM-PN, e.g. Bauer, Sterba & Hallfors, 2008). Recently, two alternative structural equation modeling (SEM) approaches for analyzing partially nested data were introduced: a multivariate single-level SEM (SSEMPN) and a multiple-arm multilevel SEM (MSEM-PN) (Sterba, Preacher, Forehand, Hardcastle, Cole & Compas, in press). In this talk, I compare and contrast these approaches and show how SSEM-PN and MSEM-PN can produce results equivalent to existing MLM-PNs. I also describe how they can be extended to flexibly accommodate several modeling features that are difficult or impossible to incorporate in MLM-PN. Importantly, implementation of such features for partially nested designs differs from fully nested designs. An empirical example involving a partially nested depression intervention combines several of these features in an analysis of interest for treatment-outcome studies.

The "Moving Anchor" for DIF detection - A Promising Approach in Need of Improvement

Paper Presentation

Carolin Strobl, *University of Zurich UZH*
Julia Kopf, *University of Zurich UZH*

When the aim is to test an item for differential item functioning (DIF) between two groups, one or more other items usually serve as an anchor for fixing the two scales. However, anchoring on particular items could also be considered as a special case of a more general restriction principle allowing any weighted sum of items to be equal between groups. Graphically this corresponds to letting the two scales move past each other - hence the name "moving anchor". This talk suggests the "moving anchor" as a new, intuitive and insightful way of thinking about anchoring. We have implemented it in a systematic grid search and present illustrations of our first results, but also ask the audience for suggestions on this work in progress, because so far we have not been able to find a criterion with which the "moving anchor" could outperform its strongest "ordinary anchor" competitors.

On the Validity and Reliability of Vignette Experiments: A Case Study on Measuring the Perceived Gender Income Gap

Poster Presentation

Dan Su, *University of Wisconsin, Madison*

Vignette experiments, also called factorial surveys, are frequently used for investigating social judgments in specific contexts. A vignette is a description of a hypothetical scenario generated from a set of context factors. In the vignette experiment, sets of vignettes are randomly assigned to respondents according to an experimental design. In order to demonstrate how vignette studies can increase the validity and reliability of measurements and effect estimates, we implement a vignette experiment on the perception of the actual and fair gender income gap in Austria that involved about 900 respondents. The vignette experiment employed a randomized block confounded factorial design with the gender gap as a between-subjects factor and the respondent's actual and fair income as anchoring vignettes. While the confounded factorial design ensures a valid and unconfounded measurement of the perceived actual and fair income, the anchoring vignettes help in increasing the measurement reliability but also in removing potential confounding in the between-subjects factor. Because the vignette experiment results in multiple measurements per respondent, we analyze the data using (i) an analysis of variance model with respondent random effects and (ii) a multilevel model.

Comparing the Performance of Item Selection Procedures in Multidimensional Computerized Adaptive Testing with Varying Test Lengths

Paper Presentation

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

The construction of assessments in computerized adaptive testing (CAT) usually involves fulfilling a large number of non-statistical constraints, such as item exposure control and content balancing. To improve measurement precision, test security, and test validity, the multidimensional priority index (MPI; Yao, 2011, 2012, & 2013) can be used to monitor many constraints simultaneously in multidimensional computerized adaptive testing (MCAT) for a stopping rule of fixed length. Because the precisions were different at different examinee levels and it resulted in a high misclassification rate, stopping rules of precision were implemented with the MPI method to maintain the same level of precision for all examinees (Yao, 2013). However, Yao's MPI method was developed for the between-item multidimensional framework. Many educational and psychological tests are constructed under a multidimensional framework. Some of the items (multidimensional items) in a test are often intended to assess multiple latent traits. Thus, a modified MPI method was proposed by the author (2013, 2014) in the within-item multidimensional CATs. In this study, two MCAT selection procedures with varying test lengths would be examined and compared through simulations.

Functional Generalized Structured Component Analysis

Dissertation Award Address

Hye Won Suk, *Arizona State University*

Generalized Structured Component Analysis (GSCA; Hwang & Takane, 2004) is a component-based structural equation modeling that aims to examine directional relationships among multiple sets of responses by combining data reduction with path modeling. In this study, we propose an extension of GSCA, called functional GSCA, to deal with functional data such as neuroimage data and physiological data, which are considered to arise from an underlying smooth curve varying over time. GSCA is limited to deal with functional data since it cannot represent infinite-dimensional smooth curves and it ignores the characteristic of functional data that the responses at adjacent time points tend to be linked. Functional GSCA aims to resolve these issues. By integrating GSCA with spline basis function expansions, functional GSCA enables to represent infinite-dimensional curves onto a finite dimensional space spanned by basis functions. An alternating penalized least squares estimation procedure is developed for parameter estimation that takes into account the smooth nature of functional data. The usefulness of the proposed method is illustrated by analyzing real data sets.

Comparison of Estimation Bias in Factor Analysis Using Bayesian and Frequentist Approaches

Poster Presentation

Yinghao Sun, *The Ohio State University*
Michael C. Edwards, *The Ohio State University*

Muthen and Asparouhov (2012) proposed a Bayesian approach to factor analysis which was claimed to have the advantages of, for example, allowing small cross-loadings to be estimated through informative priors. This study compared the estimation bias of a simple factor analysis model using Bayesian and Frequentist approaches through simulations. In the Bayesian approach, informative priors were imposed on cross-loadings, whereas in the Frequentist approach cross-loadings were fixed to zero. Considering the fact that any simple factor analysis model is unlikely to capture the true data-generating process exactly and to take into account potential population level model misfit, data were simulated from a noise-added population covariance matrix constructed using Cudeck and Browne (1992)'s method such that it had a prespecified discrepancy with the original population covariance matrix while model parameters remained unchanged. Biases in estimating inter-factor correlations and major loadings were compared across conditions with different degrees of population level discrepancy and different generating-values of major loadings, cross loadings, and inter-factor correlations. Results have shown that the Bayesian approach tended to underestimate inter-factor correlations more severely in the cases of larger discrepancy, lower major loadings, and higher inter-factor correlations, while both approaches tended to overestimate inter-factor correlations when cross-loadings were large.

Power to Detect Intervention Effects on Ensembles of Social Networks

Paper Presentation

Tracy Sweet, *University of Maryland*

The hierarchical network model (HNM) is a framework introduced by Sweet, Thomas & Junker, (2013) and Sweet, Thomas & Junker (2014) for modeling interventions and other covariate effects on ensembles of social networks, such as what would be found in randomized controlled trials in education research. In this paper, we develop calculations for the power to detect an intervention effect using the hierarchical latent space model (HLSM), an important subfamily of HNMs. We derive basic convergence results and asymptotic bounds on power, showing that standard error for the treatment effect is inversely proportional to the product of the number of ties and the number of networks; a result rather different from the usual effect of cluster size in hierarchical linear models, for example. We explore these results with a simulation study and suggest a tentative approach to power for practical applications.

Combining Propensity and Prognostic Scores in Clustered Settings

Paper Presentation

Christopher M. Swoboda, *University of Cincinnati*
Ben Kelcey, *University of Cincinnati*

Alongside propensity scores, prognostic scores have been developed to model the relationship between covariates and potential outcomes (Hansen, 2008). Prognostic scores can reduce both bias and variance of treatment effect estimators by constraining variation in the outcome due to sources other than treatment. Combining prognostic scores and propensity scores has been suggested, though remains understudied. In this study, the use of both prognostic scores and propensity scores is proposed for a multilevel setting through stratification on prognostic scores, then matching on propensity scores within stratum. Using this method, preliminary research has shown improvements in bias reduction. This approach is demonstrated in a multilevel analysis on the effectiveness of a prekindergarten program using the Early Childhood Longitudinal Study - Kindergarten Cohort data (ECLS-K).

R Package for Estimating the Multilevel Spatial Error Model

Paper Presentation

An-Shun Tai, *National Tsing Hwa University*
Pei-Hua Chen, *National Chiao Tung University*

The spatial autoregressive model has been widely used in spatial statistics and economics. Existing multilevel spatial autoregressive models studied the spatial dependence relationship at level two. Previous studies in social network data such as Add Health have used a spatial autoregressive model to study peer effects in academic achievements. However, most of these data were collected in multilevel structures and violated the assumption of between group independence. This study intends to develop an R package to estimate the multilevel spatial autoregressive model with an autoregressive error lag term at level one. Two existing R packages will be combined and linked. Two sets of simulated data will also be generated for cross validation, including different types of weight matrices. Practical implications of the proposed model will also be discussed.

The Statistical Significance of Dominance Analysis Measures in Multiple Regression

Poster Presentation

Shuwen Tang, *University of Wisconsin-Milwaukee*
Razia Azen, *University of Wisconsin-Milwaukee*

In a multiple regression model, researchers are often interested in comparing the predictors in terms of their contributions to the overall predictive effect. For this purpose, dominance analysis (Azen & Budescu, 2003; Budescu, 1993) examines the predictor's contribution by itself and in the presence of all possible combinations of the remaining predictors. The predictors can then be compared based on their dominance measures. A question researchers may further ask is whether a predictor is more important than another, over and above chance levels. The possible difficulty of making this inference is that we do not know the exact sampling distribution of the dominance measures. The technique of "bootstrapping" is frequently applied to make inferences from statistics when the sampling distribution is unknown, by sampling with replacement a large number of times from an existing data set (Efron, 1979, 1982; Efron & Tibshirani, 1986, 1993). The current study aims to adopt the bootstrapping confidence interval procedure to evaluate inferential procedures regarding a difference between the dominance measures from two predictors. A simulation study will be conducted to examine the empirical performance of the proposed procedure under different conditions. Specifically, the study will determine how the magnitude of the dominance effect size and sample size affect the performance of the proposed approach and make recommendations based on the results. This study will also include an empirical demonstration of this analysis and its appropriate interpretation.

Inflated Discrete Beta Regression Models for Likert and Discrete Rating Scale Outcomes

Paper Presentation

Cedric Taverne, *Université catholique de Louvain*
Lambert Philippe, *Université de Liège*

Discrete ordinal responses such as Likert or rating scales are regularly proposed in questionnaires and used as dependent variable in modeling. The response distribution for such scales is always discrete, with bounded support and often skewed. In addition, one particular level of the scale is frequently inflated as it cumulates respondents who invariably choose that particular level (typically the middle or one extreme of the scale) without hesitation with those who chose that alternative but might have selected a neighboring one. The inflated discrete beta regression (IDBR) model addresses those four critical characteristics that have never been taken into account simultaneously by existing models. The mean and the dispersion of rates are jointly regressed on covariates using an underlying beta distribution. The probability that choosers of the inflated level invariably make that choice is also regressed on covariates. Simulation studies suggest that the IDBR model produces more precise predictions than competing models. The ability to jointly model the location and dispersion of an ordinal response, as well as to characterize the profile of subjects selecting an "inflated" alternative are the most relevant features of the IDBR model. It is illustrated on a set of questions from the European Social Survey.

Abnormal Psychometrics: Measurement of Rare, Skewed, and Otherwise Irregular Latent Variables

Invited Speaker

Jonathan Templin, *University of Kansas*

When measuring psychological or biological disorders it is often the case that a sizeable portion of a sample may exhibit no evidence a given disorder. This type of heterogeneity is at odds with customary psychometric analyses for the development of screening instruments. More specifically, these samples do not follow the often-assumed symmetric normal or Gaussian distributions for latent variables and also do not follow a simple structure of items measuring such variables. In this talk I discuss models with non-normal latent variables, beginning with a comparison of classical normal latent variable-based psychometric models, such as factor analytic or item response models, to a newer class of non-normal latent variable diagnostic classification models. Framing the comparison as portions of a larger set of "doubly-generalized" linear models, I draw from disparate literatures: statistical mixed modeling and artificial intelligence/machine learning to show how such models have been anticipated by not widely used in psychometric research. To highlight the potential such doubly-generalized models have for informing research and clinical practice, I demonstrate the utility of differing distributions of latent variables in the analysis of a psychiatric screening instrument. I conclude the talk with a discussion of unresolved issues for psychometric analyses with doubly-generalized models.

Extending the Use of Multidimensional IRT Calibration as Projection: Many-to-one Linking and Linear Computation of Projected Scores

Paper Presentation

David Thissen, *The University of North Carolina at Chapel Hill*
Yang Liu, *The University of North Carolina at Chapel Hill*
Brooke Magnus, *The University of North Carolina at Chapel Hill*
Hally Quinn, *The University of North Carolina at Chapel Hill*

Two methods to make inferences about the scores that would have been obtained on one test using responses obtained with a different test are "scale aligning" and "projection." If both tests measure the same construct, scale aligning may be accomplished using the results of simultaneous calibration of the items from both tests with a unidimensional IRT model. If the tests measure distinct but related constructs, an alternative is the use of regression to predict scores on one test from scores on the other; when the score distribution is predicted, this is projection. Calibrated projection combines those two methods, using a multidimensional IRT (MIRT) model to simultaneously calibrate the items comprising two tests onto scales representing distinct constructs, and estimating the parameters describing the relation between the two scales. Then projection is done within the MIRT model. This presentation describes two extensions of calibrated projection: (1) the use of linear models to compute the projected scores and their error variances, and (2) projection from more than one test to a single test. The procedures are illustrated with using data for scales measuring closely related quality of life constructs.

Collusion Detection Using a Diffusion-Like Response Time Model

Paper Presentation

Anne Thissen-Roe, *Comira*
Michael Finger, *Comira*

One method for detecting test collusion, or large-scale answer sharing, is the divergence framework (Belov, 2013). It evaluates the fit of a psychometric model to the answer choices of groups of test-takers, using Kullback-Leibler divergence to identify those groups with unusual person-fit distributions. A follow-up investigation is conducted on the person-fit of individuals within anomalous groups, compared to the distribution of non-members. Another line of research depends on the identification of aberrant response times. These methods can be combined formally for greater power, using joint statistical models for item response and response time. Here, we explore the value added when collusion detection is conducted under the divergence framework, using a joint model of responses and response times extended from the diffusion family of models for choice reaction time (Ratcliff, 1978; Ratcliff, Van Zandt & McKoon, 1999).

On the Relationship between Latent Change Score Model and Autoregressive Cross-Lagged Factor Approaches and Cautions for Assessing Longitudinal Relations between Variables

Paper Presentation

Satoshi Usami, *University of Tsukuba*
Timothy Hayes, *University of Southern California*
John J McArdle, *University of Southern California*

The present research focuses on model selection problems between latent change score (LCS) and autoregressive cross-lagged factor (ARCL) models for inferring causal relationships from longitudinal designs. A large-scale simulation study is performed to investigate when discrepant results are obtained between these two models of causal inference, and what the best model selection procedure is to address this problem. Results show that differences in estimates between models are explained by the correlation between the intercept and slope factor scores in each variable and correlation between intercept scores, as well as the number of time points and sample size. Among several model selection procedures, correct specification rates are greater when examining model fit indices, compared with likelihood-ratio tests and several information criteria. An actual example using height and weight data in gerontology is also provided.

Bias and Precision in the Linear-Regression Estimate of the True Score, the Standard Errors of Measurement, and the Standard Error of Estimation

Paper Presentation

L Andries van der Ark, *University of Amsterdam*
Wilco M. H. Emons, *Tilburg University*

Within the framework of classical test theory, the unobservable true score T of a respondent having test score x can be approximated in two ways: by x itself, and by the regression of T on x (Kelley's formula) yielding , where and are the reliability and the mean of the test scores, respectively. Let denote the standard deviation of X, then under the assumptions of the classical test theory, the *standard error of measurement*   is a measure of the precision of x, and the *standard erro*r *of the estimate* is a measure of the precision of R(T|x). In finite samples, R(T|x), , and must be estimated, which requires estimates of , , and . The estimation renders two problems. First, is typically estimated by a lower bound to the reliability, such as Cronbach's alpha, so the estimates of R(T|x), , and are also biased. Second, standard errors are unavailable for the estimates of R(T|x), , and . In this study, we first derive the required standard errors. Next, we investigate how bias affects the estimates and standard errors of R(T|x), , and . Finally, we will discuss implications for the use of R(T|x), , and in small samples.

IRT Parameter Linking: Definitions, Foundational Results, and Linking Function Estimation

Paper Presentation

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

Michelle D. Barrett, *CTB/McGraw-Hill*

The problem of linking item response model parameters is due to their general lack of parameter identifiability. We characterize the formal nature of the functions required to link parameter values for the same test takers and items in different calibration studies for the general case of monotone, continuous response models, derive their specific shapes for the 3PL model, show how to identify these functions from the parameter values of common items or persons in different linking designs, and present an estimator that appears to have better behavior than the estimators of the mean/mean, mean/sigma, and Stocking-Lord linking functions currently in use.

The influence of multiple imputation on the Tucker2 model

Paper Presentation

Joost van Ginkel, *Leiden University*

Pieter Kroonenberg, *Leiden University*

Three-way analysis is an extension of Principal Component Analysis to three-way data. In three-way data usually cases (first way) are measured on several variables (second way) on different occasions (third way). The Tucker2 model is a specific three-way model in which the levels of both the first way and third way are reduced to a smaller number of components. As in any other statistical technique, missing data may both complicate the execution and interpretation of the Tucker2 model. A widely used method for dealing with missing data is multiple imputation, but relatively little has been done on multiple imputation in three-way analysis. In the current study the influence of multiple imputation on the results of the Tucker2 model is studied. Incomplete three-way data are simulated, and the missingness mechanism, the percentages of missingness, and the imputation method are varied to study their influence on the results of the Tucker2 model.

Computerized Adaptive Testing for Classifying Examinees using MIRT for Items that Measure One or Multiple Abilities

Paper Presentation

Maaike M. van Groen, *Cito\RCEC*

Theo J. H. M. Eggen, *Cito\RCEC*

Computerized adaptive tests (CATs) were developed for obtaining an efficient estimate of the examinee's ability, but they can also be used for classifying the examinee into one of two levels (e.g. master/non-master). Several methods are available for making the classification decisions for constructs modeled with a unidimensional item response theory model. These methods stop testing when enough confidence has been reached for making the decision. But if the construct is multidimensional, few classification methods are available. A classification method based on Wald's Sequential Probability Ratio Test was developed for application to CAT with a multidimensional item response theory model in which items measure multiple abilities. Seitz and Frey's (2013) method for making classifications per dimension, when items measure one dimension, was adapted for making classifications on the entire test and on parts of the test. The popular unidimensional classification method by Kingsbury and Weiss (1979), which uses the confidence interval surrounding the ability estimates, was also adapted for multidimensional decisions. Simulation studies were used to investigate the efficiency and effectiveness of the classification methods. Comparisons were made between different item selection methods, between different classification methods and between different settings for the classification methods.

Cognitive Latent Variable Models

Paper Presentation

Joachim Vandekerckhove, *University of California, Irvine*

We introduce cognitive latent variable models, a broad category of formal models that can be used to aggregate information regarding cognitive parameters across participants and tasks. Robust cognitive process models can be drawn from the cognitive science literature, and common latent structures can be coopted to combine latent ability or trait estimates across tasks. The new modeling approach allows model fitting with smaller numbers of trials per task if there are multiple participants, and is ideally suited for uncovering correlations between latent task abilities as they are expressed in experimental paradigms. An example application deals with the structure of cognitive abilities underlying executive functioning.

Item Selection in Massive Open Online Courses

Poster Presentation

Divyanshu Vats, *Rice University*
Christoph Studer, *Cornell University*
Richard Baraniuk, *Rice University*

We study the problem of selecting items for designing tests in massive open online courses (MOOCs). We assume that the responses to questions satisfy a Rasch model, where the questions are associated with a difficulty parameter and learners with an ability parameter. Standard methods for item selection are not practical since they require estimates of the learners' ability parameter, which are typically not known. By exploiting the fact that a MOOC consists of a large number of learners, we use the law of large numbers to perform efficient item selection in MOOCs when only given the mean ability parameters of all learners. This mean ability parameter corresponds to the overall ability of the learners and can be easily estimated by requiring learners to answer a small number of preliminary questions based on the prerequisites of the class. We demonstrate the effectiveness of our item-selection algorithm on real world data of over 1500 students answering 60 questions.

On the Use of Empirical and Power Priors in Bayesian Computerized Adaptive Testing

Paper Presentation

Bernard Veldkamp, *University of Twente*

Computerized adaptive testing (CAT) comes with many advantages. Unfortunately, it still is quite expensive to develop and maintain an operational CAT. In this paper, Bayesian CAT is introduced as an alternative, and the use of empirical priors is proposed for estimating item and person parameters to reduce the costs of CAT. Methods to elicit empirical priors are presented and two small examples are presented that illustrate the advantages of Bayesian CAT. Besides, special attention will be paid to power priors. These priors have been proposed to integrate information coming from historical data with current data within Bayesian parameter estimation for generalized linear models. This approach allows to use a weighted posterior distribution based on the historical study as prior distribution for the parameters in the current study. Applications can be found especially in clinical trials and survival studies. Implications of the use of empirical priors are discussed, limitations are mentioned and some suggestions for further research are formulated. The efficiency of the approach is demonstrated in terms of measurement precision by using data from the Hospital Anxiety and Depression Scale (HADS) with a small sample.

Developing a Bayesian IRT Framework for Adaptive Learning in Educational Games

Paper Presentation

Josine Verhagen, *University of Amsterdam*
D. Arena, *Kidaptive Inc.*
D. Hatfield, *Kidaptive Inc.*
S. Liu, *Kidaptive Inc.*

Educational games are becoming more and more popular, and there is an increasing demand for games that adapt to the level of the learner. In addition, is desirable to involve parents in their children’s digital learning by giving them feedback on how their learner is doing and what they can do to facilitate learning outside the digital environment. We are developing an adaptive Bayesian IRT framework to meet these two requirements simultaneously. Before the start of a game, background information about the learner and results from previous games are included in a prior for the initial ability of the learner, informing the starting level of the game. During the game, the prior ability distribution is updated after each administered item to acquire an updated ability estimate which guides item selection. The final ability distribution at the end of the game can be used to report back to parents. By including strong prior information, even a game with only a few items can be made as informative as possible. We will discuss challenges and practical problems surrounding the implementation of this framework, including processing speed on mobile devices, detection of distraction, multiple attempts at solving an item, and multidimensionality.

Using the Biadditive Model in the Context of Reliability Assessment

Paper Presentation

Jay Verkuilen, *CUNY Graduate Center*
Emily Ho, *Fordham University*

Reliability assessment makes use of data in a two-way layout without replication. The biadditive model (Denis & Gower, 1994) decomposes data of this form into additive main effects and a multiplicative interaction model that is equivalent to the singular value decomposition applied to the additive model's residuals. This model represents a general specification test for the suitability of Cronbach's alpha as a reliability coefficient. It also provides graphical diagnostics in the form of a biplot. Unlike factor analysis, it does not impose a particular functional form for the interaction and can very helpfully diagnose issues such as the presence of uniform or non-uniform DIF in a manner analogous to Stout's (2002) methodology. Even if the ultimate goal is to fit a factor model, the approach is helpful to check for model violations. It is useful in smaller sample situations such as pilot or feasibility studies and can be computed using standard statistical packages or even spreadsheets with widely available plug-ins. This makes the approach particularly helpful for users who are not psychometric specialists. We also consider the use of penalized estimation, e.g., ridge regression, to deal with estimator bias created by short scales.

Designing Simulation-& Game-Based Assessments

Paper Presentation

Alina A. von Davier, *ETS*

In this presentation I will give an overview of simulation-and game-based assessments (S&GBAs) and the status quo of their supportive research. Then I will discuss the design of two versions of a science assessment that include a simulated task and a collaborative problem solving task, respectively. I will present the data collection design, statistical and technological considerations, given that the data were collected using crowdsourcing via amazonturk. The preliminary data analyses are presented. The modeling strategies of the individual science skills and collaborative skills using both the process & outcome data are also discussed. I will also briefly present the GlassLab work on SimCityEdu and on MineCraft.

Producing Big Databases Before the 'Big Data' Buzz

Paper Presentation

Matthias von Davier, *ETS*

Some international assessments accept large amounts of missing data when imputing plausible values of proficiency variables. Currently discussed modeling approaches aim to improve the approaches taken for primary and secondary analyses while broadening the access to more policy relevant variables collected in questionnaires at different system levels. The talk will present some examples from recent cycles of large scale assessments and will provide some empirical evidence of how to improve the stability of trend measurement of cognitive skills in PISA and other assessments.

Uneducated Guesses: Three Examples of How Mistreating Missing Data Yields Misguided Educational Policy

2013 Career Award Keynote Address

Howard Wainer, *National Board of Medical Examiners*

In this talk I will describe three educational proposals that only make sense if you say them fast. In each case, their validity relies strongly on the data that are missing to have a particular structure, and in each case, that assumed missing-data structure can be dismissed. The three proposals that will be examined are: 1) make college entrance exams optional, 2) allow students to choose which test items they will answer, and 3) evaluate teachers on the gains their students show in test scores.

Equating Item Difficulty Statistics: Robustness of Poststratification versus Linear Equating Methods

Paper Presentation

Michael Walker, *College Board*

Usama Ali, *Educational Testing Service*

The precision of the item difficulty estimates is essential to get comparable item statistics and meet the test specifications for better assessment. This paper concerns two methods currently in use for equating observed item difficulty statistics. The first method involves the linear equating of item statistics in an observed sample to reference statistics on the same items. The second method, or the item response curve (IRC) method involves the summation of conditional observed item statistics across the reference population total score frequencies (i.e., poststratification).To our knowledge, there is no reference that compares the behavior of the two methods in the context of equating item difficulty indices. Therefore, using simulation studies, the paper evaluates these methods in terms of their standard error and robustness of equated item statistics using different number of studied items and sample size. Implications of the results are discussed.

Assessment of Dimensionality Can Be Distorted When Many People Have All Incorrect Answers: An Example from Psychiatry and a Solution using Mixture Models

Poster Presentation

Melanie Wall, *Columbia University*

Irini Moustaki, *London School of Economics*

Common methods for determining the number of latent dimensions underlying an item set include eigenvalue analysis and examination of fit statistics for factor analysis models with varying number of factors. Given a set of dichotomous items, we will demonstrate that these empirical assessments of dimensionality are likely to underestimate the number of dimensions when there is a preponderance of individuals in the sample with all zeros as their responses, i.e. all incorrect answers. A simulated data experiment is conducted to demonstrate this phenomena. An example is shown from psychiatry assessing the dimensionality of a social anxiety disorder battery where only one latent dimension is found if the full sample is used, while three latent dimensions are found if the excess zeroes are accounted for correctly. A mixture model, i.e. hybrid latent class latent factor model, is used to assess the dimensionality of the underlying subgroup corresponding to those who come from the part of the population with some measurable trait. Implications of the findings are discussed, in particular regarding the potential for different findings in community versus patient populations.

Asymptotic Efficiency of the Pseudo-Maximum Likelihood Estimator in Multi-group Factor Models with Pooled Data

Paper Presentation

Fan Wallentin, *Uppsala University*
ShaoBo Jin, *Uppsala University*
Anders Christoffersson, *Uppsala University*

In practice, the presence of different strata is typically unknown; pooling observations from several normal populations is an example. Distribution of pooled data becomes a mixture of normal distributions. In this study, the effect of pooling data is investigated through a two-group factor model. Two independent normal populations are pooled together. A single-group factor model is fitted to the pooled data set using pseudo-maximum likelihood (PML) where the data are treated as normally distributed and the normal theory ML is applied. The asymptotic standard errors of factor loadings for the single-group factor model are computed and compared with the asymptotic standard errors from the multi-group ML approach. Theoretically, the multi-group ML estimators should be asymptotically efficient. However, the results from our numerical study show that the PML is more efficient than the multi-group ML. A mathematical rationale shows that the standard errors from the PML are underestimated. Such underestimation is due to the ignorance of the effects of factor means and covariances in different groups. Therefore, the normal theory ML is not robust for pooled data. Especially, it largely underestimates the variances of factor loadings when error variances are larger and the group size is small.

Multivariate Hypothesis Testing Methods for Measurement of Individual Change

Paper Presentation

Chun Wang, *University of Minnesota*
David J Weiss, *University of Minnesota*

The measurement of individual change has been an important topic in both education and psychology. For instance, teachers are interested in whether students have significantly improved (e.g., learned) from instruction, and counselors are often interested in whether particular behaviors have been significantly changed after certain interventions. Recent approaches for measuring change using item response theory (IRT) include the linear logistic model (Fischer, 1989) and a multidimensional Rasch model (Embretson, 1991), among others. These approaches focused on measurement models to capture longitudinal response data. Non-IRT methods include latent growth curve modeling. However, all prior methods are concerned with testing whether growth is significant at the group level. The present research targets a novel research question: is the change in latent trait estimates for each individual significant across occasions? Kingsbury & Weiss (1983), and Finkelman, Weiss, & Kang (2010) have addressed this research question assuming the latent trait is unidimensional. We generalize their earlier work and propose three hypothesis testing methods to evaluate individual change on multiple latent traits. They are a multivariate chi-square test, a multivariate likelihood ratio test, and a Kullback-Leibler test. Simulation results show that these tests hold promise for detecting individual change with low Type I error and high power.

The Sensitivity of Hierarchical Linear Models to Outliers

Paper Presentation

Jue Wang, *University of Georgia*
Zhenqiu (Laura) Lu, *University of Georgia*
Allan S. Cohen*, University of Georgia*

The hierarchical linear model (HLM) has become popular in behavioral research, and has been widely used in various educational studies in recent years. Violations of model assumptions can have a non-ignorable impact on model. One issue in this regard is the sensitivity of HLM to outliers.

The purpose of this study is to evaluate the sensitivity of a 2-level HLM to outliers by exploring the influence of outliers on parameter estimates under the normality assumptions at both levels. Initial work using a simulation study approach was conducted using the SAS software for the random-coefficients regression model. This model is a subtype of the general HLM. We examined the differences of estimates from different models with 3 types of outliers (5 SD, 10 SD, and 15 SD), 10 levels of outlier percentages (from 5% to 50% of the whole dataset with an increase of 5%), and 2 levels of sample sizes (1250, 5000). A robust method - "Huber sandwich estimator" implemented in the SAS software, was examined and compared with the full maximum likelihood estimation method (FMLE) for dealing with level-1 outliers. The biases of the estimates varied differently, and the biases with FMLE were pretty similar to the robust method.

Dissaggregating Between-and Within-person Effects with Longitudinal Data Using Multilevel Models

Paper Presentation

Lijuan Wang, *University of Notre Dame*
Scott E. Maxwell, *University of Notre Dame*

In this talk, we extend current discussion of how to disaggregate between-person and within-person effects with longitudinal data using multilevel models. Our main focus is on the two issues of centering and detrending. We will first present some conceptual and analytical work to demonstrate the similarities and differences among three centering approaches (no centering, grand-mean centering, and person-mean centering) and the relations among various detrending approaches (no detrending, detrending X only, detrending Y only, and detrending both X and Y). Then, two real data analysis examples in psychology as well as a simulation study will be provided to illustrate the differences in the results of using different centering and detrending methods for the disaggregation of between- and within-person effects. Recommendations of how to do centering, whether detrending is needed or not, and how to do detrending if needed will be made and discussed.

Model Misspecification in Cognitive Diagnosis: Asymptotic Behavior of Maximum Likelihood Classification and a Robust Alternative

Paper Presentation

Shiyu Wang, *University of Illinois at Urbana Champaign*
Jeff Douglas, *University of Illinois at Urbana Champaign*

The maximum likelihood classification rule is a standard method to classify the examinee attribute profile in cognitive diagnosis models. Its asymptotic behavior is well understood when the model is assumed to be correct, but has not been explored in misspecified latent class models. Conditions are derived under which the MLE is consistent under model misspecification, as well as those that can lead to inconsistency. The robust DINA MLE is developed to overcome the inconsistency problem for a wide range of misspecified models. Theorems are presented to show it is a consistent estimator if the true model satisfies some general regularity conditions. Simulation studies are conducted to compare the performance of the MLE, the robust DINA MLE and a nonparametric classifier under model misspeicification. Despite providing an example in which the MLE is inconsistent, simulation results demonstrate that the MLE works remarkably well when the true model is any conjunctive model and is misspecified as a DINA model. The MLE computed under the misspecified DINA even shows empirical evidence of consistency when the true model is compensatory. These results suggest that some degree of bias can be tolerated by fitting simple and interpretable models without severely affecting classification accuracy.

Score-based Tests of Measurement Invariance: Use in Practice

Paper Presentation

Ting Wang, *University of Missouri*
Edgar C. Merkle, *University of Missouri*
Achim Zeileis, *Universitat Innsbruck*

In this presentation, we study a family of recently-proposed measurement invariance tests that are based on the *scores* of a fitted model. This family can be used to test for measurement invariance w.r.t. a continuous auxiliary variable, without pre-specification of subgroups. Moreover, the family can be used when one wishes to test for measurement invariance w.r.t. an ordinal auxiliary variable, yielding test statistics that are sensitive to violations that are monotonically related to the ordinal variable (and less sensitive to nonmonotonic violations). After providing an overview of the family of tests, we present new simulation-based results examining (i) the extent to which the tests can identify specific parameters violating measurement invariance, and (ii) the extent to which the tests are impacted by model misspecification. We also describe an R implementation that allows users to generally apply the tests to SEMs estimated under multivariate normality.

Modeling Selection Effects in Examinee-Selected Items

Paper Presentation

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

In an examinee-selected-item (ESI) design, examinees are required to respond to a fixed number of items from a set of given items (e.g., responding to 2 items from 5 given items; leading to 10 selection patterns). The ESI design has the advantages of enhancing students' learning motivation and reducing their testing anxiety. However, these advantages come at a price: scores from different combinations of items are not directly comparable. This ESI design yields incomplete data, which may be missing not at random so that standard IRT models become inappropriate. Recently, Wang et al. (2012) proposed the examinee-selected-item item response theory (IRT) model by adding an additional latent trait to account for such a selection effect. This latent trait could correlate with the target (intended-to-be-measured) latent trait. Along this research line, we developed a general class of IRT models, which include the examinee-selected-item IRT model as a special case. In the most general case, each selection pattern has one random effect to account for its distinct selection effect. Simulation results indicated a good parameter recovery for the new models. We also conducted an experiment to collect real data, in which 462 fifth graders took five pairs of mathematic (dichotomous) items. In each pair of items, students were first asked to indicate which item they prefer to answer and then answer both items. This is referred to as the "Choose one, Answer all" approach. The new IRT models were fit to the real data and the results were discussed.

Statistical Assessment of Equating Transformations

Paper Presentation

Marie Wiberg, *Umeå University, Sweden*
Jorge Gonzalez, *Pontificia Universidad Católica de Chile*

Equating methods are statistical tools that use an appropriate transformation function to map the scores of one test form into the scale of another. The equating literature shows that equating evaluation criteria might differ depending on the adopted framework. For instance, to assess equating under traditional methods, the standard error of equating, the difference that matters, and an examination of the equated scores against the raw scores are three possibilities that have been used. Within the kernel equating framework, the percent relative error is another measure that can be added to the previous list. All these measures target different parts of the equating process and aim to evaluate it from different aspects. In this paper the equating transformation is viewed as a standard statistical estimator and from that viewpoint discusses how it should be assessed. For the kernel equating framework, a numerical illustration shows the benefits of viewing the equating transformation as a statistical estimator as opposed to using equating specific criteria. A discussion on how this approach can be used to compare other equating estimators (from different frameworks e.g. IRT equating) using statistical assessment instead of equating specific evaluations is also included.

A Comparison of Measurement Concepts Across Physical Science and Social Science Domains: Instrument Design, Calibration, and Measurement

Paper Presentation

Mark Wilson, *University of California, Berkeley*
Luca Mari, *School of Industrial Engineering, Universita Carlo Cattaneo-LIUC, Castellanza (VA), Italy*
Andrew Maul, *University of Colorado, Boulder*
David Torres Irribarra, *University of California, Berkeley*

Is there a framework common to measurement in both physical and social sciences? We address this question, as its answer would determine to an important extent the possibility of building a shared measurement-related body of knowledge across these traditionally separate domains. We outline a framework of the processes involved in measurement that includes instrument design, instrument calibration, and ultimately, measurement using the instrument. A comparison of these steps across the two domains reveals both (a) formal parallelism, and (b) important differences in the way calibration is implemented. We examine the similarities and differences to determine whether they reveal a case of irreducible difference, or whether the similarities are such that measurement in the two domains can be viewed within a single conceptualization.

A General Method for Equalizing the Conditional Standard Error of Measurement (CSEM)

Poster Presentation

David Woodruff, *ACT*
Dongmei Li, *ACT*
Tony Thompson, *ACT*
Hongling Wang, *ACT*

It is usually the case in educational measurement that observed test scores are transformed to scale scores before being reported to examinees. The CSEM of observed test scores usually varies across the score scale. Scale score interpretation can be enhanced when the observed score to scale score transformation equalizes the scale score CSEM across the scale. Up until now the only method for accomplishing that equalization has been the arsine transformation in combination with the binomial error model. A more general method for finding a function of observed scores that equalizes the CSEM of the scale scores is proposed. The method may be used with both dichotomous item scores and polytomous item scores. The method uses the delta method for computing the variances of functions of random variables, and requires numerical integration. The method is illustrated for several different item score models: binomial, multinomial, and IRT. The method may be used in combination with any item score model and any procedure for finding the CSEM.

The Influence of Item Residual Heterogeneity on DIF Testing

Paper Presentation

Carol Woods, *University of Kansas*
Jared Harpole, *University of Kansas*

A limitation of logistic regression for comparing groups is that the residual variance is fixed and presumed equal for both groups. If this homogeneity assumption is violated, group comparisons are misleading. Although this limitation of logistic regression has been extensively addressed in literature on logistic regression generally, it is virtually absent from literature specifically about differential item functioning (DIF). Because DIF testing is a group comparison, DIF testing with logistic regression also will produce misleading results if the item residual homogeneity assumption is violated. Other DIF methods that use the logistic function are also likely to be biased in the presence of this type of heterogeneity. A simulation study will be presented in which Type I error and statistical power are evaluated in conditions with and without item residual heterogeneity for binary logistic regression, two-group item response modeling (IRT-LR-DIF), and the Mantel-Haenszel test.

Issues with Latent Variable Modeling: A Simulation Study

Paper Presentation

Hailemichael Metiku Worku, *Leiden University*
Mark de Rooij, *Leiden University*
Willem J. Heiser, *Leiden University*

Psychologists often collect multivariate binary data. For example, in the Netherlands Study of Depression and Anxiety (NESDA), data on depression and anxiety disorders (MDD: major depressive disorder; GAD: generalized anxiety disorder; DYST: dysthymia; SP: social phobia; and PD: panic disorder) are collected to study their prevalence and how these disorders are influenced by personality traits (Penninx et al., 2008; Spinhoven, De Rooij, Heiser, Penninx, & Smit, 2009). For the analysis of these type of data structural equation models (SEMs) are often used (Muthen, 1978; Kenneth, 1989; Skrondal & Rabe-Hesketh, 2004). Underlying the five disorders, for example, it is assumed that two latent dimensions do exist, i.e., Fear and Distress. These dimensions (factors), in turn, are dependent on personality traits. Such theories are tested by fitting SEM on the manifest variables. For binary data, tetrachoric correlations are used as an integral part of analysis in SEM. That means, an underlying continuous normal distribution for the dichotomous variables is assumed. The latent variables are also assumed to follow a normal distribution. These distributional assumptions are difficult to validate, and the validity of the model can be questionable if the assumptions are wrong. Another issue with SEM is the indeterminacy of factor scores, that is the existence of different factor scores for the same observed (indicator) variables (Gutman, 1955). We conducted a simulation study to investigate the influence of the normality assumption, sample size, number of indicator variables per factor, correlation among latent variables, communality, and whether the indicators are rare events, on the recovery of factor scores. The recovery was measured using Pearson correlation, and the occurrence of Heywood cases (Kenneth, 1989, pp. 282) and nonconvergence solutions.

Model Error as a Random Effect

State of the Art talk

Hao Wu, *Boston College*

Models are rarely exactly correct in the population and the consequence of an incorrect model is misspecification. This misspecification is usually considered as a fixed value in the population and estimated as an effect size measure of model fit. This talk will present an alternative approach where model error is assumed to arise as a random quantity and modeled with a distribution. The dispersion of this random effect serves as a measure of the size of model error. Parameters are estimated through maximum marginal likelihood and its sampling distribution accounts for the effects of both sampling error and model error. Other recent developments in the evaluation of model error will also be surveyed.

Testing Measurement Invariance for Bifactor Models

Paper Presentation

Xin Xin, *University of North Texas*

Bifactor structures are often found in educational and psychological measurement, where each item indicates one general factor and one of the group factors. Bifactor models reveal the extent of unidimensionality of data (Reise, 2012). Measurement invariance (MI) is used to establish whether the constructs are measured identically across groups. Group comparisons become valid only after MI is established. By distinguishing item response variances due to general or group factors, bifactor models could help test MI on both general and group factor levels. Additionally bifactor models may help detect systematic bias of items that may otherwise be attributed to measurement error. The present paper articulates the procedure of testing MI in bifactor models using data from TIMSS survey of 8th graders' Math attitudes. All the 12 items load on the general factor Math attitudes, and on one of the three group factors: Confidence, Affect, and Value.

A Mixture Hierarchical Model for Response Times and Response Accuracy

Paper Presentation

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

In real testing, examinees manifest one of two types of test-taking behaviors---solution behavior and rapid guessing behavior. Rapid guessing usually happens in high-stakes tests when there is insufficient time, and in low-stakes tests when there is lack-of-effort. These two qualitatively different test-taking behaviors, if ignored, will lead to violation of the local independence assumption and as a result, yield biased item/person parameter estimation. Several mixture models have been proposed to account for differences among item responses and response time (RT) patterns arising from these two behaviors. Unlike the existing mixture models which either use only response information, or RT information, or simply assume examinees only engage in one type of behavior throughout the entire test, we propose an innovative mixture hierarchical model that overcomes the limitations of currently available models. This new model uses both responses and RT information to capture the phenomenon of examinees switching between two test taking behaviors. A Monte Carlo Expectation Maximization (MCEM) algorithm is proposed for model calibration. A simulation study showed that all model parameters can be precisely recovered, and the new model fitted a real, high-stakes test data better than a non-mixture model.

Comparing Methods for Detecting Differential Item Functioning in Testlet-Based Items

Poster Presentation

Wei Xu, *University of Florida*
Anne Corinne Huggins, *University of Florida*

Differential item functioning (DIF) has drawn scholarly attention and researchers have developed statistical models to detect DIF in testlet based items. In this study, the Rasch testlet model (Wang and Wilson, 2005), the two-level testlet response model (Beretvas and Walker , 2012), the three-level multilevel measurement model (Jiao et al. 2005) and the extended logistic regression model (Sedivy, 2009) are fit to simulated data and compared with respect to their ability to detect DIF in dichotomous items nested in testlets. Three DIF magnitude (0.2, 0.4, 0.6), three testlet variances (0.5, 1, 2) and two sample sizes (500 or 2000) were manipulated. This study advances the current research by evaluating robustness of proposed models in terms of DIF detection in testlet based items.

A Diagnostic Classification Model with Multiple Response Strategies

Paper Presentation

Ning Yan, *Independent Consultant*
Yuehmei Chien, *Pearson*
Chingwei David Shin, *Pearson*

Diagnostic classification models are psychometric models that aim to diagnose the dichotomous status (mastery or non-mastery) of multiple latent attributes underlying responses on test items. A characteristic feature of such models is a binary incidence matrix, with rows corresponding to items and columns corresponding to attributes, which encodes a priori assumptions about relationships between items and attributes. The use of an incidence matrix in this form implicitly excludes the possibility that the correct answer to an item could be obtained through multiple response strategies each making use of a different combination of multiple attributes. This is particularly true when the model makes the conjunctive assumption that a correct latent response on an item depends on mastery of all the attributes specified in the corresponding row of the incidence matrix, effectively implying a unique strategy. While disjunctive models can be seen as allowing multiple strategies per item, they can only accommodate single-attribute strategies. We propose a modeling framework based on a new form of incidence matrix with columns corresponding to strategies instead of attributes. The model is conjunctive between multiple attributes within a strategy and disjunctive between multiple strategies for an item, with more flexible assumptions on success probabilities.

Handling Measurement Error in Predictors with a Multilevel Latent Variable Plausible Values Approach

Paper Presentation

Ji Seung Yang, *University of Maryland*
Michael Seltzer, *University of California - Los Angeles*

To deal with measurement error and sampling error in predictors more properly, nonlinear multilevel latent variable modeling has been suggested as an alternative to traditional multilevel modeling for situations in which latent predictors are measured by categorical manifest variables (e.g. Rabe-Hesketh et al., 2004). The purpose of this study is to propose the multilevel latent variable plausible value approach (e.g., Mislevy et al, 1992) as a relatively more proper method to handle measurement error issues in predictors in multilevel modeling settings compared to three other approaches: 1) the traditional summed score analysis, 2) traditional item response theory scale score analysis that ignores the nesting structure and uses point estimates of latent trait levels in the model for the outcomes, and 3) a latent variable plausible values approach that does not reflect nesting structure and/or that omits key covariates. As an illustrative example, we analyze the effect of a teacher's reading instructional practices on students' reading achievement using Early Childhood Longitudinal Study – Kindergarten (ECLS-K). For drawing plausible values from posterior distribution of latent teacher reading instructional practices, Markov Chain Monte Carlo (MCMC) method with Gibbs Sampling is used.

Modeling Concurrent, Cumulative, and Prospective Effects in Ecological Momentary Assessment Data

Paper Presentation

Hsiu-Ting Yu, *McGill University*

Ecological Momentary Assessment (EMA) is a reliable and valid research methodology in social science and clinical research. The methodology samples subjects' current behaviors and experiences repeatedly in real time and in subjects' natural environments. EMA data permit more sensitive assessments and reduce retrospective recall bias. Many empirical applications have focused on the concurrent effects describing the relationships between outcome and explanatory variables. The rich information of EMA data also allows researchers to examine other types of effects. This presentation shows how multilevel modeling techniques are used to model concurrent, cumulative, and prospective effects in EMA data. While concurrent effects examine the current relationships between outcome and explanatory variables, cumulative effects assess the effects of explanatory variables accumulated up to the current time point. Prospective effects indicate the predictability of explanatory variables on future outcome. Empirical data studying the relationships between ruminative behavior and negative affect are analyzed to illustrate how to model the three types of effects using multilevel modeling techniques.

The Effects of Threshold Structures on Detection of Differential Item Functioning (DIF) : A Comparison of MACS and MIMIC

Poster Presentation

Soocheol Yun, *SungKyunKwan University*
Soonmook Lee, *SungKyunKwan University*

In the present study, we compared two approaches that can be implemented to detect uniform Differential Item Functioning (DIF) in the SEM framework: the Mean and Covariance Structures (MACS) model and the Multiple-Indicator Multiple-Cause (MIMIC) model. Although both approaches are based on the same logic, assumptions and procedures for detecting DIF are not equivalent, especially when indicators are at ordinal-categorical scales. Specifically, the MACS model tells us how an item functions differently when thresholds vary across groups, whereas the MIMIC model focuses on the difference in intercepts. In other words, the difference in thresholds is assumed to be constant across groups in MIMIC models, whereas thresholds are estimated freely across groups in the MACS model. Because threshold structures dictate distribution of ordinal-categorical items, performance of the two models may differ according to difference in the threshold structures. We conducted a simulation study to compare the performance of the two models for detecting DIF under various conditions, including number of DIF items, sample size combination, threshold structure. The results and recommendations will be presented along with an emphasis on differences between two models.

Path Diagrams: Model Types, Layout Algorithms, and Some Subtleties

Paper Presentation

Yiu-Fai Yung, *SAS Institute Inc.*

Path diagrams are indispensable presentation tools in practical structural equation modeling (SEM). They not only are intuitively appealing representations of modeling ideas, but they also can be used as graphical summary of research results. Some practical researchers prefer to specify their models by drawing path diagrams in the SEM software interface, while others specify their models by writing computer code but might still want to obtain nice-looking path diagram output for presenting their statistical results. This paper focuses on the latter scenario---that is, how one can produce quality path diagrams automatically from syntactic model input. The process-flow, grouped-flow, and GRIP layout algorithms for path diagrams are described and discussed. Different layout algorithms are suitable for different types of models. Simple rules are proposed to select the best layout algorithm for a given model. Data examples are used to show how these algorithms work. Some subtleties in applying these algorithms are also demonstrated. You can expect to see a lot of path diagrams!

Rotating Spatio-Temporal Models for fMRI Time Series

Paper Presentation

Guangjian Zhang, *University of Notre Dame*

Functional magnetic resonance imaging (fMRI) time series collects blood-oxygen-level dependent (BOLD) signals of multiple brain regions over a period of time. Spatio-temporal models extract brain networks from fMRI time series. A brain network is indicated by a spatial map and the corresponding time course. Spatial maps and time courses change with different rotation methods. Influences of rotation on spatial maps and time courses of brain networks are investigated using a simulation study.

A Comparison of Equating Results for Passage-Based Tests Using Various IRT Models and Calibration Software Programs

Paper Presentation

Mengyao Zhang, *The University of Iowa*
Euijin Lim, *The University of Iowa*
Shichao Wang, *The University of Iowa*
Kyung Yong Kim, *The University of Iowa*
Hyung Jin Kim, *The University of Iowa*
Won-Chan Lee, *The University of Iowa*
Robert L. Brennan, *The University of Iowa*

The purpose of this study is to empirically investigate the effect of using various item response theory (IRT) models and calibration programs on equating results for passage-based tests. Such tests tend to involve varying degrees of local item dependence (LID), which might pose a substantial impact on IRT equating. Three perspectives of handling the LID and corresponding models are examined: (a) ignore the LID and apply dichotomous unidimensional IRT models (two- and three-parameter logistic models); (b) remove the LID by creating passage scores and use polytomous unidimensional IRT models (graded response model and generalized partial credit model); and (c) incorporate the LID by fitting multidimensional IRT models (bifactor multidimensional IRT model and testlet response model). Several commercial or open-source programs are used: BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), PARSCALE (Muraki & Bock, 2003), flexMIRT (Cai, 2012), ICL (Hanson, 2002), and R (R Development Core Team, 2005). Based on data from a national large-scale testing program, results are compared regarding test characteristic curves, test information functions, and estimated true score and observed score equating relationships. Results of this study could provide practitioners with useful information on potential discrepancies among results provided by various IRT models and software programs.

Detection of Aberrant Responses in Automated Essay Scoring

Paper Presentation

Mo Zhang, *Educational Testing Service*
Xinhao Wang, *Educational Testing Service*
Florian Lorenz, *Educational Testing Service*
Jing Chen, *Educational Testing Service*

As automated essay scoring gains in popularity in educational testing, there are many measurement issues that have not been fully resolved, one of which is the detection of aberrant responses. We define 'aberrant responses' as submissions that are either not properly automatically scorable or whose scores are unduly influenced by construct-irrelevant variation. Some aberrant responses are due to known causes (e.g., gaming-the-system, providing nonsensical responses) while others are due to heretofore unknown causes (e.g., providing responses with atypical linguistic profiles).To this date, it has not been clear which kinds of response characteristics would most likely trigger problems for automated scoring, potentially creating sources of construct irrelevant variance. Depending on the type of scoring model, such construct-irrelevant variance could be masked by acceptable machine-human agreement because such agreement typically leaves ample room for the intrusion of other factors. The goal of this study is to determine the effectiveness of various types of machine advisories in identifying aberrant responses, and to develop models to predict human-machine discrepancies. Achieving these goals will allow us to build a better understanding of the kinds of responses that may not be appropriate for automated scoring.

Methods for Moderation Analysis with Missing Data in the Predictors

Paper Presentation

Qian (Jackie) Zhang, *University of Notre Dame*
Lijuan (Peggy) Wang, *University of Notre Dame*

The most widely used statistical model for conducting moderation analysis is the moderated multiple regression (MMR) model. While conducting moderation analysis using MMR models, missing data could pose a challenge, mainly because of the nonlinear interaction term. In the study, we consider a simple MMR model, where the effect of predictor X on the outcome Y is moderated by a moderator U. The primary interest is to find ways of estimating and testing the moderation effect with the existence of missing data in the predictor X. We mainly focus on cases when X is missing completely at random, missing at random, or missing depending on auxiliary variables (missing not at random). Four methods are compared: (1) Listwise deletion; (2) Normal-distribution-based maximum likelihood estimation (NML); (3) Normal-distribution-based multiple imputation (NMI); and (4) Bayesian estimation (BE). Results from simulation studies show that the proposed methods had different relative performance depending on various factors. The factors are missing data mechanisms, roles of variables responsible for missingness, population moderation effect sizes, sample sizes, missing data proportions, and distributions of predictor X. Influences of adding auxiliary variables were also discussed in terms of estimation accuracy for NML, NMI, and BE.

Predictive Data Mining Modeling With Model Ensemble: A Demonstration and Comparison Using Large Scale Family Science Applications

Poster Presentation

Qun Zhang, *University of Kentucky*

The paper demonstrates the effectiveness of predictive data mining models in discovering meaningful information from large data sets. To that end, the paper chooses three types of predictive data mining models (decision tree, neural network, and regression), and apply them separately to two applications from family sciences where the data sets are large enough to be partitioned into separate sub-portions for effective model training, validation, and testing. One data set includes a numerical outcome whereas the other a categorical outcome. To further improve the predictions in terms of stability and/or accuracy, two model ensemble techniques (bagging and boosting) are implemented during the modeling process to pool predictions from individual component models. For comparison purposes, all models are also fitted without creating any model ensemble. Besides, the models are each evaluated for goodness-of-fit and performance at the final stage using various procedures including information criteria, ROC curve, classification table, etc. The entire analysis is performed using SAS Enterprise Miner 7.1 which provides model ensemble as well as the three types of predictive data mining models discussed here. The paper serves as the foundation for a future research topic which adds feature selection to predictive data mining modeling under model ensemble for analyzing very large data sets.

Mediation Analysis with Missing Data through Multiple Imputation and Bootstrap

Paper Presentation

Zhiyong Zhang, *University of Notre Dame*
Lijuan Wang, *University of Notre Dame*
Xin Tong, *University of Notre Dame*

A method using multiple imputation and bootstrap for dealing with missing data in mediation analysis is introduced and implemented in SAS. Through simulation studies, it is shown that the method performs well for both MCAR and MAR data without and with auxiliary variables. It is also shown that the method works equally well for MNAR data if auxiliary variables related to missingness are included. The application of the method is demonstrated through the analysis of a subset of data from the National Longitudinal Survey of Youth.

Termination Rules for the Variable-Length Cognitive Diagnostic CAT: The Standard Error Approach

Paper Presentation

Chanjin Zheng, *University of Illinois at Urbana-Champaign*
Hua-hua Chang, *University of Illinois at Urbana-Champaign*

CD-CAT purports to combine the strengths of both CAT and cognitive diagnosis (CD). CD models attempt to classify examinees into the correct latent class to produce diagnostic information whereas CAT algorithms choose items to achieve that goal as efficiently as possible. Most of the existing work adopts a fixed-length rule for terminating CAT. However, a variable-length termination rule is more desirable in order to tap the full potential of CAT in cognitive diagnosis. Two rules have been proposed by Tatsuoka (2002) and Hsu and his associates (2013). Both rules can be labeled as limited-information approaches since both only involve partial information on the cognitive pattern estimation. The current study attempts to develop two new termination rules from the standard error perspective: the attribute marginal distribution standard error rule (MDSE) and the pattern separation standard error rule (PSSE). The advantages of the new rules are two-fold: first, both of them make use of all the information on the cognitive pattern estimation. Second, they are conceptually convenient since the standard error approach is one of the dominating approaches in CAT. The preliminary results indicate that MDSE and PSSE perform better or equally as well as the previous methods.

Hierarchical Mixed Membership Stochastic Blockmodels for Modeling Network-Level Effects on Subgroup Structure

Paper Presentation

Qiwen Zheng, *University of Maryland*
Tracy Sweet, *University of Maryland*

The hierarchical network modeling framework (HNM; Sweet et al 2013) extends single network social network models for use with the naturally occurring multiple networks found in education studies in which multiple classrooms or schools are involved. We introduce a new HNM, a type of hierarchical mixed membership stochastic blockmodel (Sweet et al 2014), in which network level covariates influence how isolated subgroups are within each network. To demonstrate model fitting, we use elementary school friendship network data to determine the effects of teacher demographics on student friendship cliques.

Using Networks in Representing and Analyzing Process Data for Educational Assessment

Paper Presentation

Mengxiao Zhu, *Educational Testing Service*
Zhan Shu, *Educational Testing Service*
Alina von Davier, *Educational Testing Service*

New technology enables interactive and adaptive scenario-based tasks to be adopted in educational measurement with the benefit of high validity, reliability and quality. At the same time, it is a challenging problem to build appropriate psychometric models to analyze data collected from these tasks due to the complexity of the data. This study focuses on process data collected from scenario-based tasks (SBT). We explore the potential of using concepts and methods from social network analysis (SNA) to represent and analyze process data. Empirical data was collected from the large-scale scenario-based, computer-administered assessment of Technology and Engineering Literacy (TEL) conducted as part of the National Assessment of Educational Progress (NAEP) project. For the activities sequences in process data, directed networks are created with nodes representing actions and directed links connecting two actions only if the first action is followed by the second action in the sequences. This study shows initial results of using SNA in visualizing process data and displaying the problem solving processes. This study also explores the potential of using the ideas from exponential random graph models (ERGMs) in identifying meaningful patterns in process data. Finally, this study discusses the insights these analyses can provide on task/item design.

Using Fixed Parameter Calibration for Stable Item Parameter Estimation under Less Ideal Pretesting Conditions for a Large-scale Assessment

Paper Presentation

Rongchun Zhu, *ACT, Inc.*
Xiaohong Gao, *ACT, Inc.*

It is often required to pretest a large volume of items at the beginning stage of a large-scale test program, for example, new tests under preparation by state consortia of common core. Stable item parameter estimates are also needed for both pre-equating fixed forms and computerized adaptive tests so that high-quality score reports can be immediately produced to inform subsequent instructional or other assessment decisions. However, due to various factors, for example, lack of test-taker motivation, pretest samples tend to perform worse than prospective test population, which makes it difficult to obtain stable item parameter estimates based on pretesting data alone. This study investigates various settings of fixed parameter calibration to better calibrate items when an IRT model is used. Initial operational samples are proposed to be used as a baseline to evaluate pretest samples. Based on such findings, more reasonable specifications are applied in calibration with various methods and models also explored. Related practical implications will be discussed.

Estimation of Mixed Models for Polychotomous Data Using Auxiliary Variables

Paper Presentation

Bonne J.H. Zijlstra, *University of Amsterdam*

For the analysis of generalized linear mixed models, Bayesian simulation methods are sometimes preferred because they allow prior information to be included. Generally, these simulation methods can be expected to be most efficient if sequential sampling from the conditional distributions of the model parameters can be applied (Gibbs sampling). However, for the models at hand some conditional distributions are intractable and therefore need to be replaced by less efficient accept-reject steps. In mixed models for dichotomous data, the application of some of these steps can be prevented by introducing auxiliary latent variables. It will be investigated whether this approach is also applicable to models for polychotomous data and to what extend this increases the computational efficiency compared to conventional Bayesian simulation methods.

NOTES

SPONSORS

ACT®

CollegeBoard

Mc Graw Hill Education | CTB

海云天科技
SEA SKY LAND

IEA

ETS®
Listening. Learning. Leading.®

GMAC®
GRADUATE MANAGEMENT
ADMISSION COUNCIL

eMetric

NATIONAL BOARD OF MEDICAL EXAMINERS 1915 U.S.A.
NBME®

sas®

stata®

UNIVERSITY OF
WISCONSIN SYSTEM
CENTER FOR
PLACEMENT TESTING

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Cover artwork courtesy Keary Kautzer of Sheboygan, Wisconsin