IMPS 2016

Asheville, NC, USA • July 11-15, 2016

Abstract Book: Talks

Monday, July 11, 2016

All Day Workshops: 8:00 AM - 9:30 AM

Workshop 1: Computerized Adaptive Testing and Multistage Testing with R
David Magis, University of Liège, Belgium; Duanli Yan, Educational Testing Service, USA; Alina A. von Davier, Educational Testing Service, USA

Workshop 2: flexMIRT®: Flexible Multilevel Multidimensional Item Analysis and Test Scoring
RJ Wirth, Vector Psychometric Group, LLC; Carrie R. Houts, Vector Psychometrics Group, LLC

Workshop 3: Bayesian Modeling Using
Ben Goodrich, Columbia University; Daniel Furr, University of California at Berkeley

Tuesday, July 12, 2016

Keynote: 8:15 AM - 9:15 AM

Keynote: Simplifying the use of latent class analysis by means of stepwise modeling approaches
Jereon Vermunt, Tilburg University

Chair: Klaas Sijtsma

I will give an overview of recent and ongoing work of my group on various types of stepwise modeling approaches for LC analysis. Most of this work is part of a large research project funded by the Netherlands Science Foundation. It includes research on three types of very promising approaches: 1. The use of measures similar to modification indices for model fit assessment and model adjustment in simple and complex latent class analysis 2. Bias adjusted three-step latent class analysis for studying the relationship between class membership and external variables 3. Divisive latent class analysis for the construction of latent class trees, yielding an approach similar to hierarchical cluster analysis.

Item Response Theory-IRT 1: 9:30 AM - 11:00 AM

IRT 1a: Examine the Practical Gain of Modeling Response and Response Time
Jiyun Zu, Educational Testing Service; Frederic Robin, Educational Testing Service

Response time (RT) refers to the amount of time a test-taker spends on answering an item. A Bayesian hierarchical model for jointly modeling response and RT (van der Linden, 2007) uses RT as collateral information in estimating the response parameters (e.g., item parameters and person ability). Item responses and RT are modeled separately at the first level, and are integrated at the second level through jointly modeling the random effects item and person parameters. This approach has been shown to reduce the bias and standard error of item and ability parameter estimates, particularly when the correlation of person ability and speed is high (van der Linden, Klein Entink & Fox, 2010).

Molenaar, Tuerlinckx & van der Maas (2015) simplified the Bayesian hierarchical model to a generalized linear factor model (GFM) by treating the item parameters as fixed and showed it did not affect item parameter recovery under realistic circumstances. However, they did not discuss the quality of ability estimates, which could be the reported scores.

The main purpose of this study is to examine the practical gain of jointly modeling response and RT using data from two testing programs, respectively measuring English language listening skill, and quantitative reasoning. By conducting exploratory analyses on RT, modeling responses only, and modeling responses and RT using the Bayesian hierarchical model and GFM, we study the effect of modeling RT on the point and precision of item and ability parameter estimates, and compare results from the Bayesian hierarchical and GFM approaches.

IRT 1b: Psychometric Analysis of Situational Judgment Tests

Youn Seon Lim, University of Illinois at Urbana-Champaign; Fritz Drasgow, University of Illinois at Urbana-Champaign

The Situational Judgment Test (SJT) is used for assessing job candidates' social competence and problem solving skills in the workplace. Traditional approaches to study the latent structure underlying SJT items rely on factor-analytic methods. This study concerns an alternative evaluation of SJTs based on the theory of cognitive diagnosis. Of particular interest is whether and how factor analytic methods can be used to devise a Q-matrix for the SJT (The Q-matrix of a cognitive diagnostic test specifies the item-attribute associations of the test.) In addition to examining the performance of factor analytic methods as a tool to devise the Q-matrix, multidimensional item response theory (MIRT) models and non-parametric clustering and classification are examined. The psychometric properties of these methods are investigated in simulation studies and their comparative performance for a real data set is reported.

IRT 1c: Evaluate Item Dependency in an Aptitude Reasoning Test

Youn Seon Lim, University of Illinois at Urbana-Champaign; Fritz Drasgow, University of Illinois at Urbana-Champaign

The Special Tertiary Admissions Test (STAT), developed by the Australian Council for Educational Research (ACER), is as an alternate method of gaining entry into Australian university courses, for people who do not hold a recent Year 12 certificate in Australia. The test compromises 64 scored multiple-choice items, half of which are Verbal Reasoning questions and the other half, Quantitative Reasoning questions. Items were designed in testlets, where they could share the same stimulus (example, text paragraph).

To investigate the level of item dependency within each testlet, two methods were implemented and compared for 2014/15 STAT data, comprising more than 8,000 candidates, 55% of which were female and 45% of which were male. The first popular method was computing residual correlations using RUMM software (Andrich, Sheridan, & Luo, 2012). Residual correlations are often used to determine if any pair of items shows a sign of dependency. The other method was using item bundle fit statistics from ConQuest software (Adams, Wu, & Wilson, 2011) for each testlet. Additional simulation data sets with similar structure to STAT were also generated to compare the two dependency detecting methods.

Initial results show that items in both domains fitted well to the Rasch model. The level of item dependency in each of the testlets was low. Additionally, the study suggests that item bundle fit statistics by ConQuest could overcome a limitation of the residual correlations for detecting independency among a set of three or more items rather than two.

IRT 1d: Mixture Model with Internal Restrictions on Item Difficulty

Evan Olson, University of Maryland; Hong Jiao, University of Maryland

In the model with internal restrictions on item difficulty (MIRID; Butter, De Boeck & Verhelst, 1998) a composite item's difficulty parameter is dependent on the estimated item difficulty of its components. For assessments where these composite and component items are used, a beneficial utility of the MIRID is that the regression weights of the components can be estimated, and are not required to be established a priori as in the linear logistic test model (LLTM; Fischer, 1973). This study examines a proposed Mixture MIRID (MixMIRID) in an extension of the MIRID that incorporates item parameter estimation by latent subgroups. These subgroups are defined by the finite mixture distribution as determined from the data. A simulation will be performed using Markov chain Monte Carlo estimation to explore study conditions with varied test lengths, mixing proportions, and numbers of latent subgroups. The application to a real data set will also be conducted to demonstrate the proposed model.

Factor Analysis- FAC 1: 9:30 AM - 11:00 AM

FAC 1a: Random Coefficient Confirmatory Factor Analysis Models with Bayesian Lasso Prior

Junhao Pan, Department of Psychology, Sun Yat-sen University, ChinaLaurette Dubé, McGill Univeristy

Multivariate repeated measures data, in which the measurement occasions are nested within individual or subject, are widely used in the social sciences. In this paper, under a hierarchical Bayesian framework, we proposed a general random coefficient confirmatory factor analysis (RC-CFA) model with LASSO prior. Besides the theoretical and practical advantages provided by RC-CFA for modeling heterogeneity, the LASSO prior was assigned to the entire residual covariance matrix of the observed indicators for each individual. Therefore, different from the usually diagonal assumption, the residual covariance matrix can be modeled as a sparse positive definite matrix that contains only a few off-diagonal elements bounded away from zero. Markov Chain Monte Carlo (MCMC) procedures were developed to perform Bayesian inference. The Bayesian method achieves model parsimony and generally fits the data better, while keeping the factor structure intact. In other words, the number of factors and the form of factor loading matrix remain unchanged for different individuals. Both simulated and real data sets were analyzed to evaluate the validity and practical usefulness of the proposed procedure.

FAC 1b: Accuracy of Bi-Factor Exploratory Rotations with Orthogonal Structures

Luis Garrido, Universidad Iberoamericana; Eduardo García-Garzón, Universidad Autónoma de Madrid; Juan Barrada, Universidad de Zaragoza; Julio Olea, Universidad Autónoma de Madrid; Francisco Abad, Universidad Autónoma de Madrid

Despite the dramatic increase in the popularity of bi-factor models in the last decade, little is known about the accuracy of the different rotation methods that are available to uncover bi-factor structures in an exploratory context. The current study aims to address this issue by examining the performance of five rotation methods in the recovery of unrestricted orthogonal bi-factor structures: Schmid-Leiman (SL), Target Rotation based on SL (SLt), Iterated Target based on SL (SLi), Bi-Quartimin (BFq) and Bi-Geomin (BFg). Using Monte Carlo methods, several salient variables were systematically manipulated, including sample size, factor loadings on the general and specific factors, presence of cross-loadings, presence of pure indicators of the general factor, and number of variables per specific factor. The results revealed that SLi accurately recovered the bi-factor structures across the majority of the conditions, and generally outperformed the other methods. As predicted by theory, SL showed inferior levels of accuracy when pure indicators of the general factor were present, while BFq and BFg poorly recovered the bi-factor structures in the presence and absence of cross-loadings, respectively. Also, although SLt seemed to improve the overall performance of SL, in some instances such as without cross-loadings it tended to worsen it. Further results suggest that under especially adverse conditions, such as with very small samples or low loadings for the specific factors, a poor recovery of the bi-factor structures is expected regardless of rotation method. Based on these findings, the authors offer practical guidelines and provide examples with empirical data.

FAC 1c: Estimation Approaches for Modeling Categorical Indicators with Low Endorsement

Sierra Bainter, University of Miami

Generalized linear factor analysis (GLFA) is a general latent variable modeling framework that can be applied to continuous or categorical measures, subsuming traditional psychometric factor analysis and item response theory models. A recurring issue for estimating GLFA models with categorical indicators is low item endorsement (item sparseness), due to limited sample sizes or extreme items such as rare symptoms or behaviors. Researchers cannot always avoid sparse items, and many psychological constructs necessarily involve studying items with low endorsement (e.g., risky behaviors, illicit drug use). In this presentation, I demonstrate that under conditions characterized by sparseness, currently available estimation methods, including maximum likelihood (ML), are likely to fail. Even if models converge, ML estimation is likely to lead to extreme estimates and low empirical power. Bayesian estimation is a promising alternative to ML estimation for GLFA models with item sparseness. Results from a simulation study demonstrate the advantages of Bayesian estimation to stabilize estimates, aid convergence, and improve empirical power.

FAC 1d: On Examining Specificity in Latent Construct Indicators

Tenko Raykov, Michigan State University

A latent variable modeling procedure for examining specificity in any indicator of a common factor for a given set of measures is outlined in a longitudinal design setting with measurement invariance and specificity stability over time. The method permits one to test whether there is specificity in a given factor indicator, and in the affirmative case allows one to point and interval estimate the specificity variance in that measure. The procedure aims to contribute to clearing possible confusion regarding indicator specificity in some applied and methodological factor analysis literature. The discussed method is readily applicable with popular software, is based on empirically testable conditions, and is illustrated using a numerical example.

Estimation and Computational Methods-ECM 1: 9:30 AM - 11:00 AM

ECM 1a: Perceived Beneficial Modeling - A More Interpretable Analysis than R-Squared

Johnson C. Li, University of Manitoba

Pearson's correlation coefficient (r) is widely employed to evaluate the linear relationship between two interval or ratio variables (e.g., intelligence X and SAT scores Y). The squared correlation (r^2) is also frequently reported to quantify the effect size, i.e., the proportion of variance of Y (e.g., SAT scores) that is accounted for by X (e.g., intelligence). Despite its popularity, r^2 is usually regarded as a complicated statistical concept and many researchers and practitioners may misinterpret its meaning (Brooks, Dalal, & Nolan, 2014; Dunlop, 1994). In light of this, I am proposing and developing an appealing new framework—Perceived Beneficial Modeling (PBM)—which is based on previous research on stochastic dominance statistics (e.g., Cliff, 1993, Ruscio, 2008). Under PBM, I have developed a statistical estimate (Ar) that describes the X-Y relationship in a more intuitive way as compared to "the proportion of variance explained" provided by r^2. For example, with intelligence and SAT scores, Ar = .667 means that a person who is above-mean in intelligence has a 66.7% likelihood of scoring above-mean in SAT score. The likelihood or perceived benefit of 66.7% should be more easily understood and interpreted than the conventional r^2 = .25. Simulation results of the parametric and non-parametric Ar estimates and the associated confidence intervals will be presented. Furthermore, potential and implication of PBM in more complex relationships than the simple bivariate X-Y relationship (e.g., regression with covariates, mediation, and moderation) will be discussed.

ECM 1b: Robust Confidence Intervals for Standardized Regression Coefficients

Paul Dudgeon, University of Melbourne

Yuan and Chan (2011) rigorously derived consistent confidence intervals for standardized regression coefficients under general conditions. They demonstrated that large-sample confidence intervals for unstandardized regression coefficients could not in general be simply rescaled and applied to standardized coefficients when normality assumptions hold. Jones and Waller (2015) extended these development to circumstances where data are non-normal by applying Browne's (1984) asymptotic distribution-free (ADF) theory. This talk proposes heteroscedastic-consistent (HC) estimators as potentially better solutions for constructing confidence intervals on standardized regression coefficients under non-normality. A new method for estimating HC standard errors based on misspecified structural equation modelling is proposed (Yuan & Hayashi, 2006). Both the ADF and HC estimators are evaluated in a Monte Carlo simulation. Findings confirm the superiority of a subset of HC estimators over the ADF estimator and over Yuan and

Chan's normal theory method. Possible extensions of the HC estimator method to other effect sizes in linear regression and to other modelling contexts will be considered.

ECM 1c: Robustness of Several Effect-Size Measurement Methods Under Non-Normal Likert-Point Items.

Yongtian Cheng, University of Manitoba; Johnson C. Li, University of Manitoba

This Monte Carlo study evaluated the robustness of common effect size (ES) estimates (Cohen's d, probability-based A, point-biserial correlation rpb) and robust Cohen's d (dR) (Algina, Keselman & Renfield, 2005) under non-normally distributed Likert-data. ES estimates, d, rpb, dR, Rpb, and A, quantify the strength of difference between two groups of observations. However, there's little empirical evidence regarding their robustness when the scores are measured on a Likert-point scale, which is frequently used in psychology. Several factors were examined: population d (.2,.5,.8), ratio of standard deviation for two groups (.25,1,4), sample size (25,50,100), and number of Likert points (5,7). For the 5-point and 7-point items, we categorized the scores based on Beal and Dawson's (2007) method. The dataset obtained from this method has similar characteristics to Rusico and Roche's (2012) real-world non-normal dataset. This design produced 234 conditions. Each condition was replicated 5,000 times producing 1,170,000 simulated data-sets.A was found to be the most robust. In the 5-point condition, the mean absolute percentage error for A was 2.68%, d was 11.28% rpb was 60.82% and dR was 15.46%. In the 7-point condition, the mean percentage error for A was 2.02%, d was 11.85%, rpb was 72.39% and dR was 6.76%. The implications of the study are that under non-normal Likert data, A is more robust than others, which is the least influenced by the non-normality of dataset, and that the most-common used d is sensitive to the variance rate. d may not be used under Likert data that has high variance ratio. (e.g., clinical sample).

Applications-APP 1: 9:30 AM - 11:00 AM

APP 1a: Evaluating Standard Setting Methods for Higher Education

Samantha Bouwmeester, Erasmus University Rotterdam

There is an ongoing debate about the evaluation of the psychometric quality of the exams at the psychology institute of the Erasmus University Rotterdam. One point of discussion is the standard setting method that is used to define levels of achievement or proficiency and the cut scores corresponding to those levels. In the currently used methodology an absolute (criterion) reference is used that requires 55% correct responses after correction for guessing to pass an exam. A disadvantage of this methodology is that it does not take into account fluctuations in the difficulty and the reliability of the exams. In order to evaluate which standard setting method leads to the fewest errors in fail/pass decisions for individual students we conducted a simulation study. We simulated true scores and observed scores obtained by five methods to correct for guessing and several standard setting methods like the Hofstee method (Hofstee, 1983) and (variations of) the Cohen-Schotanus (Cohen-Schotanus and colleagues, 1996). We evaluated these methods for a broad range of realistic exam situations varying among others: a) length of the test; b) number of alternatives; c) discrimination of the test items; and d) difficulty of the test. The results are still preliminary by now but they already show considerable differences between different guessing correction methods and standard setting methods. In the presentation these differences will be discussed as well as the influence of the factors on the performance of the different methods.

APP 1b: Supporting Diagnostic Inferences Using Significance Tests for Subtest Scores

William Lorié, Pearson USA

Users of content-heterogeneous assessments based on unidimensional trait models often call for more information about examinee strengths and weaknesses in specific subareas. This is commonly called diagnostic information, and a standard way of providing it is to compute and report subscores. However, in the many cases where subscores fail to provide reliable information sufficiently independent of the total score (Sinharay, 2010), they cannot support claims about subarea strengths and weaknesses relative to total score expectations. These claims are referred to here as diagnostic inferences. This talk introduces a method to support diagnostic inferences for test programs developed and maintained using Item Response Theory (IRT). The method establishes null and alternative hypotheses for the number correct on subsets of items, or subtests. Statistical significance testing is then conducted to determine the strength of the statistical evidence in favor of a particular diagnostic inference. If the subtest score is modeled as a Poisson binomial distribution with probabilities set to those expected by the IRT model conditional on fixed item parameters and person scores, then testing for diagnostic inferences can be conducted for individuals or groups. The talk will present the results of power computations showing the subtest lengths generally required for supporting diagnostic inferences under different conditions and effect sizes. The relationship between traditionally reported subscores and diagnostic inference flagging is explored with data from an adaptively delivered statewide grade 6 mathematics assessment.

APP 1c: Prediction of Student Performance: Secondary Analysis using Smart PLS

Lihua Xu, University of Central Florida; Debra McGann, University of Central Florida

This study was a secondary analysis of NAEP 2008 8th-grade visual arts data (n = 3912). The purpose was to predict students' visual art response performance using students' home environment, personal characteristics, school curriculum and extracurricular activities. Formative measurement models and structural paths were modelled in structural equation modeling (SEM) using Smart PLS. The initial SEM model included four latent constructs and one endogenous variable measuring students' performance. Both direct and indirect effects between latent constructs were modelled and assessed. Number of books at home and parental education level were identified as the most important predictors of Home Environment. "I like to do artwork" and "People tell me I am a good artist" were identified as the most salient predictors of Personal Attributes. "Do you keep an art journal or sketchbook in school?" and "Art class: paint or draw?" define the School Curriculum the best. "Not for school: make artwork" and "Not for school: keep an art journal or sketchbook" define the Extracurricular Activities the best. Altogether, the four latent constructs explained 21.3% of the variance in students' responding performance, with Home construct displaying the strongest impact. Surprisingly, as compared to home environment, students' personal attributes and their art-related extracurricular activities predict students' performance to a substantially lesser degree. School-related painting and drawing and other artistic activities in school do predict students' performance significantly and yet in a lesser strength. Implications of these findings will be discussed.

APP 1d: Latent Variable Hybrid Modeling in Type D Personality

Robert Hillen, Tilburg University

This study explores the use of latent variable hybrid models for studying dimensional and categorical properties of latent variable structures. Using distressed (Type D) personality as an example, we discuss the conceptual underpinnings of hybrid modeling, illustrate the steps in building hybrid models, and explain how to decide between competing hybrid models. Three different latent variable hybrid models were fitted to distressed personality data from a large (N = 1,587) sample from the general Dutch population, each representing different conceptions of Type D personality. These conceptions include a purely dimensional view, a hierarchical view including first-order dimensions of NA and SI and higher order Type D categories, and a dimensional view with an interacting Type D construct. The model with two latent classes fitted best, but inspection of the classes showed small between-class differences compared to the within-class variability. The dominance of within-class variability supports a dimensional view on Type D personality. A follow-up simulation study on the power and sensitivity of the fit indices to class separation showed that a combination of fit indices performed moderately in detecting the two-class model. Future research may focus on improved methods to gauge the within-class and between-class differences within latent-mixture modeling to aid substantive interpretation.

Structural Equation Modeling-SEM 1: 9:30 AM - 11:00 AM

SEM 1a: Model Selection in GSCA Using Confirmatory Tetrad Analysis

Ji Hoon Ryoo, University of Virginia; Heungsun Hwang, McGill University

Structural equation modeling (SEM) is used not only for testing the validity of the hypothesized model, but also for comparing several theoretically plausible models to find the best fitting one to sample data. The latter is not uncommon in the SEM literature. For example, researchers often compare several models representing different degrees of measurement and/or structural invariance. Fit indexes play a key role in such model selection, utilizing the chi-square statistic that is obtained via maximum likelihood estimation (MLE). However, MLE requires a restrictive assumption, such as multivariate normality, that may not be valid in large data. Generalized structured component analysis (GSCA; Takane & Hwang, 2004) has been considered a practical alternative for more than a decade. GSCA is based on least squares estimation, so that it does not require the normality assumption. Moreover, it can be applied to both small and large data. GSCA currently provides four fit indexes, such as FIT, AFIT, GFI, and SRMR to evaluate a model and/or compare multiple models (Hwang & Takane, 2015). However, there still are issues in using these indexes, including no provision of cutoffs for the indexes. In this study, we apply vanishing tetrads to compare models in GSCA. Confirmatory tetrads analysis (CTA; Bollen & Ting, 1993) has been used as a substitute for likelihood ratio tests for model comparison, because it allows comparing non-nested models and providing a fit index for unidentified models. We will present an application of CTA and compare the result with the four fit indexes in GSCA.

SEM 1b: lsl: A Package for Structural Equation Modeling with Penalized Likelihood

Po-Hsien Huang, National Cheng Kung University

Penalized likelihood (PL) is now a popular method for many statistical learning problems. By utilizing sparsity inducing penalties, the underlying relationship among variables can be efficiently explored under the PL framework. Recently, Huang, Chen, and Weng (2014) proposed a PL method for structural equation modeling (SEM). Their PL method can be seen as a compromise of traditional SEM and exploratory SEM. If the penalty level is well chosen, the PL method could yield a final model that balances the tradeoff between model goodness-of-fit and model complexity. In this presentation, we will demonstrate how the SEM with PL can be implemented via the R package lsl (latent structure learning; Huang, 2015). lsl requires users to specify SEM models by indicating which parameters should be fixed, estimated freely, or estimated with penalization. Through the learning process, Isl outputs a final SEM model according to Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). lsl could also print the the values of parameters under different penalty levels. Finally, future directions for improving the lsl package will be briefly discussed.

SEM 1c: Bootstrapping Confidence Intervals for Fit Indices in SEM

Xijuan Zhang, University of British Columbia

Bootstrapping approximate fit indices in structural equation modeling (SEM) is of great importance because most fit indices do not have tractable analytic distributions. Model-based bootstrap, which has been proposed to obtain the distribution of the model chi-square statistic under the null hypothesis (Bollen & Stine, 1992), is not theoretically appropriate for obtaining confidence intervals for fit indices because it assumes the null is exactly true. On the other hand, naïve bootstrap is not expected to work well for those fit indices that are based on the chi-square statistic, such as the RMSEA and the CFI, because sample noncentrality is a biased estimate of the population noncentrality. In this article we argue that a recently proposed bootstrap approach due to Yuan, Hayashi, and Yanagihara (YHY; 2007) is ideal for bootstrapping fit indices that are based on the chi-square. This method transforms the data so that the "parent" population has the population noncentrality parameter equal to the estimated noncentrality in the original sample. We conducted a simulation study to evaluate the performance of the YHY bootstrap and the naïve bootstrap for four indices: RMSEA, CFI, GFI, and SRMR. We found that for RMSEA and CFI, the confidence intervals (CIs) under the YHY bootstrap had relatively good coverage rates for all conditions whereas the CIs under the naïve bootstrap had very low coverage rates when the fitted model had large degrees of freedom. However, for GFI and SRMR, the CIs under both bootstrap methods had poor coverage rates in most conditions.

SEM 1d: Comparing Estimators for Nonlinear Structural Equation Models Under Misspecification

Nora Umbach, University of Tuebingen, Germany; Holger Brandt, University of Tuebingen; Augustin Kelava, University of Tuebingen; Kenneth Bollen, The University of North Carolina at Chapel Hill

Models developed to answer research questions in the social sciences are most probably always misspecified. Although a higher complexity of models can be incorporated in structural equation models (SEM), misspecifications such as omitted variables or cross loadings can typically be assumed. The purpose of the current simulation study is to compare different estimators for nonlinear SEMs and to explore how they perform under different conditions of misspecified models. Some research has examined the consequences of structural misspecification in linear structural equation models, but few have examined misspecification in nonlinear SEMs. In the meantime, nonlinear SEMs (e.g., including interaction effects) are often used in the social sciences. Therefore, comparing different estimators that can fit nonlinear structural SEMs has important practical implications. The two-stage least square estimator (2SLS; Bollen, 1995; Bollen & Paxton, 1998) has been shown to be more robust than the system-wide maximum likelihood estimator with respect to parameter bias for misspecified latent linear models (Bollen, Kirby, Curran, Paxton, & Chen, 2007). The current study compares the 2SLS estimator to two other estimators for nonlinear structural equation models: The two-stage method of moments estimator (2SMM; Wall & Amemiya, 2003) and the nonlinear structural equation mixture model approach (NSEMM; Kelava, Nagengast, & Brandt, 2014). A large range of different conditions (reliability of indicators, sample sizes, normality of data) is covered and several misspecifications for nonlinear SEM are investigated. Results show that the 2SLS estimator shows less bias for nonlinear SEMs than 2SMM and NSEMM. Practical implications of these results are discussed.

Measurement Invariance and Differential Item Functioning- DIF 1: 9:30 AM - 11:00 AM

DIF 1a: A More General Model for Testing Measurement Invariance and DIF

Daniel J. Bauer, The University of North Carolina at Chapel Hill

A long standing and active areas of psychometric research concerns the establishment of Measurement Invariance (MI). This focus is understandable, as we can only validly (and fairly) compare individuals or groups on measures that have the same meaning and metric for all parties. Individual difference research, therefore, depends critically on MI. At present, the dominant strategies for evaluating MI involve applying latent variable models (factor analysis or item response theory models) and testing whether the parameters relating the items to the latent traits are equivalent for individuals with different characteristics. Differences in the item parameters reflect Differential Item Functioning (DIF) and call measurement invariance into question. The two primary models used to evaluate MI/DIF are the Multiple Groups (MG) model and the Multiple-Indicator Multiple-Cause (MIMIC) model. The MG model can be used to test the equivalence of any model parameters across levels of a single predefined grouping variable. In contrast, the MIMIC model is restricted to testing the equivalence of only intercept and mean parameters but can evaluate MI/DIF with respect to both discrete and continuous characteristics. The current presentation demonstrates that the recently introduced Moderated Nonlinear Factor Analysis (MNLFA) model offers an alternative, more flexible approach for evaluating MI/DIF which subsumes and combines the strengths of the MG and MIMIC models. Specifically, it is shown that MNLFA allows for a full and simultaneous assessment of MI/DIF across multiple categorical and/or continuous individual difference variables.

DIF 1b: DIF and Factorial Invariance for Scales with Seven Response Alternatives

David Thissen, University of North Carolina at Chapel Hill

Measurement invariance, as it is known in the factor analytic tradition, and a lack of differential item functioning (DIF), as it is known in the item response theory (IRT) literature, are essentially the same conceptually. However, the two types of analysis have rarely been cast in exactly parallel form. Either categorical (IRT) or continuous linear confirmatory factor analysis (CFA) models might usefully be applied to item response data with seven (plus or minus about two) response alternatives. This presentation is intended to clarify some of the issues involved in the application of CFA to DIF detection, by using the CFA procedures slightly differently than they are used in the conventional analysis of factorial invariance as a prerequisite for measurement invariance. Capsule summaries are provided of DIF detection using parametric IRT models, and the conventional evaluation of factorial invariance. The evaluation of factorial invariance is reformulated to make it more like DIF detection, for the context of item analysis. An empirical illustration is provided using both parametric IRT DIF detection procedures and a parallel version of the evaluation of factorial invariance using CFA models.

DIF 1c: Permutation Randomization Methods for Testing Measurement Equivalence/Invariance Across Groups

Terrence D. Jorgensen, University of Amsterdam; Benjamin A. Kite, University of Kansas; Po-Yi Chen, University of Kansas; Stephen D. Short, College of Charleston, USA

In multigroup factor analysis, measurement invariance is accepted as tenable when researchers observe (a) a nonsignificant chi-squared test of exact fit to establish configural invariance and (b) nonsignificant chi-squared difference tests of metric and scalar invariance. Large samples yield high power to detect negligible misspecifications, and chi-squared is biased under common conditions (e.g., missing data, discrete indicators). Some researchers thus prefer alternative fit indices (AFIs). Fixed cutoffs have been proposed for evaluating the effect of invariance constraints on AFIs (e.g., Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008). We demonstrate that all of these cutoffs have inconsistent Type I error rates. As a solution, we propose replacing chi-squared and fixed AFI cutoffs with permutation tests. Randomly permuting group assignment results in average between-group differences of zero, so iterative permutation yields an empirical distribution of any fit measure under the null hypothesis of invariance across groups. Simulations show that the permutation test of configural invariance controls Type I error rates better than chi-squared or AFIs when the model contains parsimony error (i.e., negligible misspecification) but the factor structure is equivalent across groups (i.e., the null hypothesis is true). For testing metric and scalar invariance, Type I error rates are better controlled by permutation tests and the chi-squared difference test than by using fixed cutoffs for change in AFIs. Permutation also yields similar power to the chi-squared difference test. Extensions are proposed for controlling Type I error rates when testing multiple items for lack of invariance.

DIF 1d: Unifying DIF in Categorical CFA and GRM

Yu-Wei Chang, University of Michigan; Rung-Ching Tsai, National Taiwan Normal University; Nan-Jung Hsu, National Tsing-Hua University

The multiple-group categorical confirmatory factor analysis model (CFA) and the graded response model (GRM) are commonly used to examine polytomous items for differential item functioning to detect possible measurement bias in educational testing. In this study, the multiple-group categorical CFA and normal-ogive GRM models are unified under the common framework of discretization of a normal variant. We rigorously justify a set of identified parameters and determine possible identifiability constraints necessary to make the parameters just-identified and estimable in the common framework of multiple-group categorical CFA models (MCCFA). By doing so, the difference between categorical CFA and normal-ogive GRM is simply the use of two different sets of identifiability constraints, rather than the seeming distinction between CFA and GRM. Thus, we compare the performance on DIF assessment between the categorical CFA and GRM approaches through simulation studies on the MCCFA models with their corresponding particular sets of identifiability constraints. Our results show that, under the scenarios with varying degrees of DIF for examinees of different ability levels, models with the GRM-type of identifiability constraints generally perform better on DIF detection with a higher testing power. General guidelines regarding the choice of just-identified parameterization are also provided for practical use.

Symposium 1: Quantification Theory fromModern Perspectives: 9:30 AM - 11:00 AM

Symposium 1a: An Outcry of Dual Scaling: The Key is Duality

Pieter Kroonenberg, University of Leiden; Shizuhiko Nishisato, University of Toronto

Quantification theory, known by such names as dual scaling (DS), correspondence analysis, homogeneity analysis and optimal scaling, is a method to quantify a matrix of non-numerical or categorical data. As was stressed at the Beijing meeting last year by Nishisato, its traditional basis lies in the simultaneous treatment of rows and columns of a data matrix, in short, dual scaling, which poses a different way of mapping data in multidimensional space. The dual aspect of analysis is clearly demonstrated by the simultaneous linear regressions (Hirschfeld, 1935), the method of reciprocal averages (Horst, 1935) and the name dual scaling (Nishisato, 1980). The current paper will revisit those points discussed last year in Beijing with further illustrative examples so as to make the basic framework of dual scaling clearer than before. Therefore, we resort to intuitive understanding of the process and a number of basic known steps of quantification theory will be reiterated and illustrated in details so that we eventually be able to grasp the real intention of 'doubled multidimensional space' (Nishsato, 2014), a basis for total information analysis (TIA), or comprehensive dual scaling (CDS) (Nishisato & Clavel, 2010). The current talk will be followed by Clavel & Nishisato, who will provide substantial enough numerical examples to justify and validate the raison d'être of TIA (CDS) and proper examples of dual quantification theory in multidimensional space.

Symposium 1b: Total Information Analysis in Practice

José G. Clavel, Universidad de Murcia

Following the introduction by Nishisato, we will present the detailed formulation of total information analysis (TIA) or comprehensive dual scaling (CDS), originally proposed by Nishisato and Clavel (2010). TIA is a means to offer a satisfactory solution to the perennial problems associated with multidimensional graphical display in quantification theory, be it symmetric graph (French plot), or non-symmetric graph or a variety of biplot. In this presentation, we will start with basic formulas for TIA, specifically those for within-set and between-set distances (Nishisato and Clavel, 2003), based on all components derived from dual scaling. Then, we will look at numerical examples to illustrate computations for TIA. In applied fields of data analysis, there are many questions about the need for dual multidimensional analysis such as: 'We carried out cluster analysis of consumers (row variables) and cluster analysis of commercial products (column variables). How can we find the cluster configuration of consumers and commercial products in the same multidimensional space?'; 'We have clusters of consumers and clusters of products. How can we compare the two sets of clusters?' Another aspect of TIA is to look at the data structure of both rows and columns in total space, not in reduced space as is often done for the sake of feasibility – we are well aware that in reduced space we fail to capture, or rather ignore some rare, but idiosyncratic relations between row and column variables. TIA will capture such relations as 'total' indicates total space.

Symposium 1c: Confidence Ellipses and p-values for Correspondence Analysis

Pieter M. Kroonenberg, Leiden University & The Three-Mode Company; Eric Beh, University of Newcastle; Rosaria Lombardo, Second University of Naples

Correspondence analysis provides the analyst with an analytical tool for visualising the association structure between two or more categorical variables. The most utilised feature of the analysis is therefore the correspondence plot which helps distinguish those categories that have a similar (or different) impact on the association structure by observing their proximity from the origin. However, how close, or far, away from the origin does a point need to be for a category to play a statistically significant role in defining the association structure between the variables?

To help answer this question, the analyst may consider confidence regions such as those proposed by Lebart et al. (1984, pp. 182 – 186), Beh (2001). Computational procedures such as those described by Greenacre (2007, pp. 196 – 197) and Ringrose (2012) also provide confidence regions using bootstrapping. For relatively few, or large, number of variables the task of performing these calculations may be relatively trivial or excessive. Beh (2010) proposed a simple means of deriving confidence regions that are elliptical in shape and were further discussed in Beh & Lombardo (2014a, 2014b). They showed the three key advantages of these ellipses are 1) they reflect the different inertia values attached to each axis, 2) they reflect information contained in the higher dimensions, and 3) p-values for each category can be calculated. We shall be discussing the use of these elliptical regions and p-values in various contexts including non-symmetrical correspondence analysis, for the analysis of multiple categorical variables and for analysing ordered categorical variables.

Symposium 1d: Polynomial Biplots for Ordered Three-Way Correspondence Analysis

Rosaria Lombardo, Second University of Naples; Eric J Beh, University of Newcastle; Pieter M. Kroonenberg, Leiden University & The Three-Mode Company

The association in two-way contingency tables is often modelled and portrayed using correspondence analysis based on the singular vectors from the singular value decomposition. In the case of three-way tables three-mode correspondence analysis uses the components here also called singular vectors from a three-mode component analysis. The symmetric or asymmetric association between the three variables is then portrayed using these singular vectors in joint plots and/or interactive or nested-mode biplots (Carlier & Kroonenberg, 1996; Kroonenberg, 2008, chap. 17). When the variables are ordinal, such plots do not make use of this ordinality in displaying the association. However, the ordinality can be exploited by replacing the singular vectors by orthogonal polynomials (see, also, Beh & Davy, 1998; Lombardo, Beh, & Kroonenberg, 2015).Polynomial biplots explicitly use the orthogonal polynomials to portray trends in the total, marginal and partial association among ordinal variables in contingency tables. The specific characteristics of these polynomials can then be used to enhance the interpretation of the nature of the association. Two-way polynomial biplots was proposed earlier by Beh & Lombardo (2014, chap. 6-7) and Lombardo, Beh & Kroonenberg (2015). Here, we will concentrate on polynomial biplots for three-way tables without a dependence structure among the ordered variables. However, in the case of a dependence structure among the ordered variables, similar biplots can be constructed.

Symposium 1e: Two Causes of the Horseshoe Phenomenon in Multiple Correspondence Analysis

Takashi Murakami, Department of Sociology, Chukyo University

The existence of monotonic items with different response rates is the first cause of the horseshoe phenomenon in multiple correspondence analysis (MCA) (which is equivalent to principal component analysis) of binary data. However, another important cause lies in analysis of items with more than two response categories; p items with c ordered categories. Two-dimensional solution of the dataset by MCA consists of linear projections of many regular (c-1)-simplexes on a plane, a configuration of some apexes of p sets of mutually congruent polygons, which themselves tend to form small horseshoes. We demonstrate that most ordered categorical variables yield the horseshoe phenomenon from the second cause even if they have essentially multidimensional structure, and we propose a method for avoiding the horseshoe phenomenon by rotating the solution space. We also propose an empirically meaningful interpretation of the horseshoe as the extreme response set when it occurs in analysis of Likert-type items. Our proposals diverge at least partially from the commonly accepted view that the horseshoe phenomenon occurs from the overwhelming one-dimensionality of analyzed data and should be ignored as an artifact.

Dissertation Award Co-winners: 11:15 AM - 11:45 AM

Generalized Fiducial Inference for Graded Response Models

Yang Liu, University of California, Merced

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

Advances in Choice Response Time Modeling: Non-Linearity, Efficient Computation, and Non-Decision Time

Stijn Verdonck, KU Leuven, University of Leuven

Choice RT experiments are an invaluable tool in psychology and neuroscience. In this talk, I will give a quick overview of my doctoral dissertation, covering three key aspects of choice RT modeling: the decision model itself, efficient computation, and parameter estimation.

As fundamental modeling work, a non-linear diffusion model for choice RT is mathematically derived from the emerging collective behavior of pooled populations of stochastic binary neurons (the Ising Decision Maker or IDM). Apart from giving a parsimonious account of typical choice RT data, the IDM makes a series of non-intuitive predictions (e.g., the so-called van der Molen-Keuss effect), that were never before related to diffusion models.

Regarding computational efficiency of diffusion model simulations (necessary for the statistical inference of the more complex models), significant progress is made by exploiting the massively parallel computation capabilities of GPUs. Dedicated code is developed that allows the GPU-based simulation of a broad class of diffusion models, providing desktop platform speeds that rival those of small high performance computing clusters.

Finally, an alternative approach to the estimation of choice RT models is proposed that conveniently bypasses the specification of the non-decision time distribution by means of an unconventional convolution of data and decision model distributions (hence called the D*M approach). Once the decision model parameters have been estimated, it is possible to compute a non-parametric estimate of the non-decision time distribution. The technique is extensively tested on simulated data, and is shown to systematically remove traditional estimation bias related to misspecified non-decision time, even for a relatively small number of observations. Reanalyzing a number of existing diffusion model application papers with the D*M method resulted in substantially different parameter estimates, which in one case even lead to a radically different qualitative conclusion.

Symposium 2: Virtual International Collaboration: 1:45 PM - 3:15 PM

Symposium 2a: Virtual International Collaboration
Onder Subul

Abstract is not available at this time

Symposium 2b: Virtual International Collaboration
Remo Ostini

Abstract is not available at this time

Symposium 2c: Virtual International Collaboration
Anil Kanjee, Tshwane University of Technology

Abstract is not available at this time

Symposium 2d: Virtual International Collaboration
Ruben Klein, Cesgranrio Foundation

Abstract is not available at this time

Symposium 2e: Virtual International Collaboration
Abstract is not available at this time

Validity and Reliability-VAL 1: 1:45 PM - 3:15 PM

VAL 1a: Rater Agreement in Test-to-Curriculum Alignment Procedures: A Meta-Analysis
Anne Traynor, Purdue University

During the development of large-scale school achievement tests, recruited panels of independent Subject-Matter Experts (SMEs) use systematic judgmental methods—often collectively referred to as “alignment” methods—to rate the correspondence between a given test’s items and the objective statements in a particular curricular standards document. High disagreement among the expert panelists, which has sometimes been observed, may indicate problems with training, feedback or other steps of the alignment procedure (Davis-Becker & Buckendahl, 2013). While procedural recommendations for implementing alignment reviews appear in the applied literature, these recommendations have been derived largely from single-panel research studies; support for their use during operational large-scale test development is limited. We have obtained over 100 alignment review reports published between 2000 and 2015 for 34 US states’ Grade K-12 achievement tests by a systematic web search and contact with state assessment directors. Using multilevel meta-regression to synthesize information from the reports, we will determine to what extent specific characteristics of test-standards alignment review procedures, particularly the number and qualifications of SME reviewers, content area, grade level, year, number of test items, number of objectives in the standards document, components of and duration of training, and type of feedback given to SMEs during the procedure, appear to affect observed agreement or reliability among expert raters (i.e., intraclass correlation coefficients or proportion pairwise agreement) about the correspondence between particular test items, and objective statements in curricular standards documents.

VAL 1b: Estimating Single-Item Reliability for Real Tests and Questionnaires

Eva A.O. Zijlmans, Tilburg University; Andries van der Ark, University of Amsterdam; Jesper Tijmstra, Tilburg University; Klaas Sijtsma, Tilburg University

Item-score reliability is of interest in psychological and educational testing practice, especially in the context of selecting appropriate items during test construction that reliably distinguish people. In previous research (Zijlmans, Van der Ark, Tijmstra, & Sijtsma, 2016), we developed four methods to estimate single-item reliability. In the present study, we applied the four methods to several real-data sets for tests used in different fields of psychology and educational measurement to estimate single-item reliability for the items in each of these tests. For each of the four methods and each test, we discuss single-item reliability values we obtained, and we also discuss how researchers can use single-item reliability to assess the quality of the items and the items' contribution to the reliability of the test score. In addition, we also discuss potential problems researchers might encounter when using single-item reliability, and if possible we distinguish advantages and disadvantages for each of the four single-item reliability methods. Finally, we discuss what single-item reliability information adds to other information sources about individual items, such as item-rest correlations, item H-scalability coefficients from Mokken scale analysis, and discrimination parameters from IRT.

VAL 1c: Estimating the Accuracy and Consistency of Change Score Classifications

Joe Grochowalski, The College Board

The purpose of this study was to introduce a method for estimating the accuracy and consistency of classifications based on change scores from a test administered at two time points to the same group of test takers. Change scores are becoming increasingly important as they're used for assessing developmental benchmarks as well as for treatment (e.g., teaching, clinical) efficacy. Methods exist to estimate the classification accuracy and classification consistency (CA and CC, respectively) for individual scores, but no method has yet focused on the special case of change scores.   Change scores are composite scores, so decisions based on cut scores will depend on the accuracy of the scores that were used to make the composite (change) score. Thus, this method takes the univariate approach of Livingston & Lewis (1995) and adapts it for use with multivariate (i.e., scores at time 1 and time 2) distributions. To evaluate the utility and accuracy of the method, a simulation analysis was used to generate scenarios with different true score distributions, different score correlations between times 1 and 2, and different reliabilities at times 1 and 2. The CA and CC of the multivariate method were compared to the univariate method (i.e., the univariate difference score distribution). Results show that when the original scores from times 1 and 2 used for the composite have different distributions, the multivariate CA and CC estimates were generally more accurate than the univariate estimates.

VAL 1d: Testing Criterion Correlations with Measurement Errors Using Latent Variable Modeling

Youngjun Lee, Michigan State University; Tenko Raykov, Michigan State University; George A. Marcoulides, University of California, Santa Barbara; Siegfried Gabler, Leibniz Institute for the Social Sciences

A latent variable modeling method for testing criterion correlations with measurement error terms in multi-component measuring instruments is outlined. The approach is based on an application of the Benjamini-Hochberg multiple testing procedure and can be used when assumptions of validity estimation related procedures need to be examined. The method also allows studying the extent to which criterion validity coefficients may be due to the relationship between a presumed underlying latent construct evaluated by a psychometric scale and a criterion variable, or may be a consequence of the relation between measurement error in the overall scale score and the criterion. The discussed procedure is widely applicable with popular latent variable modeling software, and is illustrated using a numerical example.

Computer-Based Testing- CBT 1: 1:45 PM - 3:15 PM

CBT 1a: Multidimensional Test Design Approach for Ability Estimation in Adaptive Testing

Eren Halil Ozberk, The University of North Carolina at Greensboro; Elif Bengi Unsal Ozberk, The University of North Carolina at Greensboro; Terry Ackerman, The University of North Carolina at Greensboro; Richard M Luecht, University of North Carolina Gree

A test can be designed for many purposes, including the ranking of people along a continuum or providing diagnostic value about examinees. However, a very common problem that often arises is when items are capable of measuring unwanted dimensions. Unidimensional item response theory (UIRT) assumes the items are measuring the same trait or composite of multiple traits. When a test is measuring multiple dimensions one could identify items measuring the separate dimensions and apply UIRT to each set of items. Another approach would be to use multidimensional IRT (MIRT) and simultaneously calibrate both dimensions. The purpose of this study is to compare domain and composite ability estimation accuracy in multidimensional computer adaptive testing (MCAT) with respect to MIRT test design and calibration strategies. A two-dimensional item pool was simulated with simple and complex structure. Dimensions correlated at $\rho = .3, .6$, and .9. Three calibration strategies, separate unidimensional and two multidimensional (Bock and Aitkin's EM and Metropolis-Hastings Robbins-Monro algorithm) calibration were examined. The multidimensional CAT item selection procedures: minimum angle, minimize the error variance of the composite score with the optimized weight, and Kullback–Leibler (KL) information were also examined. Item parameters were generated from $a \sim LN\{0, 0.2)$, $b \sim N(0, 1)$, and $c \sim Beta\{6,16)$ distributions. For multidimensional test structure design 1000×2 matrix of true ability parameters were randomly generated from the multivariate normal distribution $N(\mu, \Sigma)$. The domain and composite ability accuracy was evaluated using Pearson's product-moment correlation and the root mean square error (RMSE).

CBT 1b: Scoring Incomplete Adaptive Tests

Qi Diao, Pacific Metrics Corporation; Wim van der Linden, Pacific Metrics Corporation; Hao Ren, Data Recognition Corporation

Adaptive testing is based on the principle of adapting the selection of the test items to updates of the examinee's ability estimate. Most current rules for scoring adaptive tests are based on the assumption that the examinee does respond to all items assigned to him by the testing engine. However, in real-world testing programs, examinees are sometimes unable to complete their tests because of a technical failure in the infrastructure during their session. The question addressed in this presentation is how to score the incomplete response patterns for such examinees. An advantage offered by the shadow-test approach (STA) to adaptive testing is permanent projection of the optimal remaining portion of the test for each examinee along with the selection of the best item from it for administration. As technical failures can be supposed to be random, the problem becomes one of scoring incomplete response vectors under the assumption of the responses for some of the known items missing at random (MAR). The proposed study investigates how to obtain the most informative ability estimate along with an estimate of the uncertainty due to nonresponse. Our statistical framework is Bayesian framework with multiple imputation to reflect sampling variability. The results will help us to answer the question of how to score incomplete adaptive tests best from a statistical point of view. In addition, policy decisions for cases where the MAR assumption does not hold (e.g., examinees terminating the prematurely test for strategic reasons) will be discussed.

CBT 1c: A New Stopping Rule for Licensure Tests

Xiao Luo, National Council of State Boards of Nursing (NCSBN); Doyoung Kim, National Council of State Boards of Nursing (NCSBN); Ada Woo, National Council of State Boards of Nursing (NCSBN)

The stopping rule which determines when to terminate a variable-length computerized adaptive test (CAT) is an essential component of CAT. A well-designed stopping rule helps a CAT achieve desirable measurement accuracy using fewest items. Two types of stopping rules have been widely used (Dodd, Koch, & De Ayala, 1993): the standard error (SE) and the minimum information (MI) rules. The SE rule terminates the CAT after the SE decreases below a certain threshold. whereas the MI rule terminates the CAT when no items provide a minimum level of information. Additionally, Choi, Grady and Dodd (2010) proposed the predicted standard error reduction (PSER) stopping rule which terminates the CAT when the administration of an additional item fails to reduce the standard error significantly.

In licensure settings where the purpose of the test is to make a pass/fail licensure decision, the CAT is typically terminated after a clear decision can be made. That is, the 95% confidence interval of the current ability estimate is clearly above or below the cut score(s). This study proposes a new stopping rule for licensure tests, which bases the stopping decision on the predicted lower- and upper-bound of the ability estimate. Specifically, the method is to find a way to predict the ability estimate/distribution when the test continues to the end as closely as it would be in actuality. We expect this method to obtain a narrower confidence interval than using local information alone and thus to be more efficient.

Symposium 3: Recent Advances in Social Network Methodology:
1:45 PM - 3:15 PM

Cross-Validation for Social Network Data

Beau Dabbs, Carnegie Mellon University

We propose a cross-validation method for social network data that allows us to perform model selection among a large class of social network models with conditionally independent dyads (CID models). This class of models includes models that take advantage of latent nodal covariates, such as the stochastic block model (SBM), mixed membership stochastic block model (MMSBM), and latent space distance model (LSM).

Most latent variable models for social network data use some function of nodal latent variables to represent the propensity for ties in a network. However, it is frequently unclear for a given network which latent variable model would be most appropriate. Further, the latent variable models we mentioned above also have dimensionality parameters that must be chosen before model parameters can be estimated. Our cross-validation method allows us to select among these models, while also determining the dimensionality parameter as well.

We analyze the performance of this method by first performing a simulation study where the true network generation process is known. We are able to show that our method is accurate in determining the true generative model, and compare the cross-validation method to other model-based approaches such as AIC and BIC. Next, we examine some of the analytic properties of cross-validation estimators. We show that cross-validation risk estimators are approximately unbiased for networks generated from Conditionally Independent Dyad models. We also consider the variance of cross-validation risk estimators for network data and present best practices for choosing folds when performing K-fold cross-validation.

Symposium 3a: Networks Modeling for Problem Solving Processes

Alina A. von Davier, Educational Testing Service, USA; Mengxiao Zhu, Educational Testing Service

In the original definition, social networks consist of humans as nodes and relations among humans as links. In the field of education research, social network analysis (SNA) was applied to model human systems, that is, the interactions among students and teachers in the classroom, or in the online settings. However, researchers from various fields, including physics, biology, computer science, and others, extended the network studies to include systems constructed among non-human entities. This presentation applies social network techniques in modeling and analyzing networks constructed from the problem solving processes in education. In some of these networks, nodes are students. In other cases, the nodes are extended beyond human beings, and represent areas of interest, actions, and skills. The links represent the connections and transitions among the nodes. This presentation introduces two recent studies in this direction and demonstrates the novel ways of applying SNA in education. In the first study, social networks were used to model students' eyeball movements while answering science items in an online-delivered assessment. Networks were constructed to represent the transitions between areas of interest (AOIs) for students during the problem solving processes. In the second study, groups of students were evaluated through discussions on scenarios related to modern engineering problems with no clear solutions. We constructed two types of networks, the communication among students and the connections among engineering professional skills during the discussion process. In both studies, we show different network patterns for high-performing individuals/groups and for low-performing individuals/groups.

Symposium 3b: Different Specifications of Reciprocity in Social Network Analysis.

In social network analysis, the pattern of relations between actors is analyzed. One pattern that often is an important factor determining the structure of social networks is the degree to which relations are reciprocated, also called reciprocity. Different models for social network analysis, however, apply different parametrizations for reciprocity. Several models will be compared with respect to reciprocity and a new alternative will be presented.

Symposium 3c: Temporal Latent Space Network Model with Covariates

Sam Adhikari, Carnegie Mellon University

In many real world applications of longitudinal social network analysis, it is important to understand how the relationship between observed nodal covariates and tie formation changes over time. For example, in instructional advice seeking network of teachers (Spillane et al., 2012), education researchers are interested in how advice seeking behavior of the teachers changes with time, given the interventions in the network at different time points. Additionally, the researchers also want to understand how similarities or differences in observed nodal covariates affect the present and the future advice seeking behavior of the teachers in the network. Inference related questions, such as the ones in longitudinal advice seeking network, is not usually the focus of existing longitudinal latent variable network models. In this talk, I will present my work on the extension of the latent space approach of Hoff et al. (2002), to account for longitudinal dependencies among networks, and introduce a Bayesian longitudinal latent space network model with covariates. We leverage on dual interpretation of the latent inter-nodal distance as latent variable and as residual to include edge covariates in the model. We then apply the proposed model to analyze longitudinal advice seeking network of teachers.

Symposium 3d: Modeling Social Networks as Mediators

Tracy Sweet, University of Maryland

Interventions on systems (e.g. schools, workplaces) sometimes aim to change the ways in which individuals interact as means to improve some other outcome. For example, educational professional development programs often focus on the ways in teachers communicate or consult with each other to ultimately improve teaching quality. In these studies, the social network within each school provides insight into the mechanisms that affect individual outcomes and acts as a power mediator between the intervention and outcome.

Sweet et al. (2013) and Sweet et al. (2014) introduced statistical network models for network-level interventions, building on a small literature on multilevel social network models. In their models, they introduce models for samples that accommodate treatment effects. These models naturally lend themselves to education research where networks of teachers (or students) exist within each school.

We introduce a Bayesian framework for modeling social networks as mediators; we combine statistical network models for network-level interventions with Bayesian mediation models (Yuan and MacKinnon, 2009). Instead of a mediating variable, we develop models where the social network is the mediator. As a proof of concept, we introduce a mediation model where the mediator is a hierarchical mixed membership stochastic blockmodel. Thus, the intervention affects the shape of the networks and the shape of the networks affects a node-level variable (e.g. teacher instruction).

Invited Speakers: Ric Luecht; Carolin Strobl: 3:30 PM - 4:15 PM

Engineering Design in the Assessment World: A New Paradigm for Test
Ric Luecht, University of North Carolina at Greensboro

Chair: Luz Bay

Psychometric theory and modeling technologies have made enormous contributions to the measurement field. However, we still seem to lack a coherent framework for reconciling our rather sophisticated psychometric models and scaling methods with test design and development practices. The latter requires serious consideration of domain-specific content and cognitive attributes in designing and writing items, and also impact the test assembly process. Assessment engineering (AE) is a relatively new framework comprised of five building blocks for concretely integrating content-focused item and test development practices with psychometric modeling and analysis. The AE building blocks borrow strong design and quality control principles from industrial engineering to provide robust and scalable assessment solutions. The ultimate goal is to create an efficient manufacturing system for producing valid, reliable, low-cost, high-quality items and performance tasks that can be adapted for use in assessment settings ranging from formative assessments in classrooms to high-stakes certification/licensure tests (e.g. potentially generating massive top-quality item banks that provide on-target measures of multiple traits or attributes classes, and that do NOT require pretesting). This presentation will discuss some of the critical issues leading to the development of AE, will lay out the five building blocks comprising the framework, and provide some real-life examples of AE-related research to date.

Model-Based Recursive Partitioning of Psychometric Models: A Data-Driven Approach for Detecting Heterogeneity in Model Parameters
Carolin Strobl, University of Zurich

Chair: Anders Skrondal

Model-based recursive partitioning is a flexible framework for detecting differences in model parameters between two or more groups of subjects. Its origins lie in machine learning, where its predecessor methods, classification and regression trees, had been introduced around the 1980s as a nonparametric regression technique. Today, after the statistical flaws of the early algorithms have been overcome, their extension to detecting heterogeneity in parametric models makes recursive partitioning methods a valuable addition to the statistical “toolbox” in various areas of application, including econometrics and psychometrics. This talk gives an overview about the rationale and statistical background of model-based recursive partitioning in general and in particular with extensions to psychometric models for paired comparisons as well as item response models. In this context, the data-driven approach of model-based recursive partitioning proves to be particularly suited for detecting violations of homogeneity or invariance, such as differential item functioning, where we usually have no a priori hypotheses about the underlying group structure.

Item Response Theory-IRT 2: 4:25 PM - 5:55 PM

IRT 2a: Estimation of Nonnormal Latent Densities in Multilevel Item Factor Models
Ji Seung Yang, University of Maryland, College Park; Jeffrey R. Harring, University of Maryland - College Park; Xiaying Zheng, University of Maryland, College Park; Ji An, University of Maryland-College Park

When the assumption of a normally distributed latent trait in population is not reasonably held in item response theory (IRT) applications, researchers can take the nonnormality into account using non-Gaussian distributions such as the empirical histogram (EH; Bock & Aitkin, 1981; Mislevy, 1984), nonparametric curves (e.g., Ramsey or Davidian curves; Woods & Thissen, 2006; Monroe, 2014), the skewed t distribution (Asparouhov & Muthén, 2015), and finite mixture of latent distributions (e.g., Cho, Cohen, & Kim, 2014).

While the nonparametric IRT models or EH approach allows the maximum flexibility by accommodating any shape of latent density, there is little guidance for researchers on which model should be chosen between the parametric vs non-parametric approaches. The purpose of this study is to provide guidelines for researches who are adjudicating which of the two approaches is most appropriate by exploring the behaviors of model fit statistics and assessing the recovery of nonnormal latent densities when non-parametric (multilevel Ramsey-curve IRT; Yang, Zheng, & An, 2015) and parametric (multilevel mixture IRT) models are used to analyze item response data collected in multilevel settings. A simulation study is conducted to cover different sampling conditions (number of clusters, cluster size, and magnitude of intra-class correlation), shapes of distributions (number of latent classes, distance among classes, and skewness), and measurement conditions (long or short tests). Both multilevel Ramsey-curve IRT and mixture IRT models are applied to simulated data, and the results are assessed with respect to model fit statistics, item parameters, and nonnormal latent densities at the cluster level.

IRT 2b: Standard Errors of Two-Level Scalability Coefficients
Andries van der Ark, University of Amsterdam; Bonne J.H. Zijlstra, University of Amsterdam; Letty Koopman, University of Amsterdam

Tests with a multilevel structure occur often in educational and psychological research, where subjects are assessed by multiple raters. For example, a classroom of students evaluating their teacher. Multilevel nonparametric item response theory models account for this structure, and assess the quality of such measurement instruments by means of within- and between-rater scalability coefficients. To enable hypothesis testing and confidence interval construction of these coefficients, we derived standard errors for the multilevel scalability coefficients. We demonstrate the derivation and show a real-data application.

IRT 2c: Comparing Hierarchical Signal Detection Theory and Polytomous Item Response Theory
William Muntean, Pearson; Joe Betts, Pearson; Doyoung Kim, National Council of State Boards of Nursing (NCSBN)

Psychological experiments often involve two-alternative forced choice recognition tests. An early methodology for measuring recognition accuracy in these types of experiments relied on a crude method of subtracting the number of false alarms (endorsing unstudied stimuli) from the number of correct hits. This obfuscates an important responding behavior: endorsement bias. For this reason, psychologists adopted a more intuitive analysis based on signal detection theory (SDT), which assumes stimuli vary across an underlying memory strength and become endorsed when passing a criterion—a minimum threshold. Accordingly, SDT derives two latent trait estimates, sensitivity and endorsement bias. A hierarchical SDT model that nests observations within a single stimuli parallels the general structure of many common polytomous item response theory models (IRT). However, unlike SDT, polytomous IRT models do not separately account for responding bias. That is, IRT jointly models ability and responding behavior. Despite this fact, they remain popular in psychometrics. The current research employs several simulations to investigate the similarities and differences between SDT and IRT. Initial data are generated from a SDT model, a partial credit model, a graded response model, and a testlet model. Then, by fitting each model to the different generated data, we clarify the relationship between the different approaches. On the one hand, relative to IRT analyses, SDT provides unique information about responding behavior. On the other hand, IRT ability estimates are highly correlated with SDT sensitivity estimates. The implications of this relationship is discussed along with the potential for using both methods for analyzing psychological data.

IRT 2d: Confidence Intervals for Intraclass Correlations in Multilevel Item Factor Models
Xiaying Zheng, University of Maryland, College Park; Ji Seung Yang, University of Maryland, College Park

Multilevel item response theory (IRT) models (e.g., Fox & Glas, 2001) have been developed to address the clustering of examinees in educational and psychological studies. The intraclass correlation coefficient (ICC) is an important parameter in multilevel IRT models that describes how strongly examinees within a cluster correlate with each other in terms of latent abilities. However, the interval estimation of the ICC estimates has been less explored compared to point estimates, particularly in multilevel IRT. One challenge for constructing the ICC confidence intervals is the lack of a closed-form solution due to unknown distribution of the ICCs. The goal of the research is to assess five approximation methods for ICC confidence intervals in multilevel IRT, including F-distribution approximation (Donner, 1979; Tomas & Hultquist, 1978), Fisher’s z transformation (Fisher, 1925), delta method (Smith, 1956; Swiger, 1964), beta-distribution approximation (Demetrashvili, Wit, & van den Heuvel, 2014) and empirical bootstrapping. These methods have not been comprehensively examined, especially when the outcome variable is latent. A simulation study is conducted to examine the five methods under various data conditions, including balance/unbalanced designs, sample sizes, magnitudes of ICCs, and distributions of the latent scores (i.e., normal/non-normal distributions). Three parameterization methods for multilevel IRT models are used to estimate the variance components and calculate the ICCs, including fixing level-1 scale, fixing level-2 scale, and fixing the parameters of one item. All models are estimated with full-information maximum likelihood. The coverage probabilities of the true ICC are compared across the approximation methods and model parameterization methods.

Psychometrika Anniversary Session 1: 4:25 PM - 5:55 PM

Psychometrika Anniversary Session 1a

Willem Heiser, Leiden University

Willem Heiser on Carroll and Chang (1970)

Psychometrika Anniversary Session 1b

Klaas Sijtsma, Tilburg University

Klaas Sijtsma on Cronbach (1951)

Psychometrika Anniversary Session 1c

Hans Friedrich Koehn, University of Illinois at Urbana-Champaign

Hans Friedrich Koehn on Greenhouse and Geisser (1959)

Psychometrika Anniversary Session 1d

Larry Hubert, University of Illinois at Champaign-Urbana

Lawrence Hubert on Akaike (1967) and Horn (1965)

Measurement Invariance and Differential Item Functioning- DIF 2: 4:25 PM - 5:55 PM

DIF 2a: Different Methods' Performances for DIF Based on the Ability Estimations

Levent Ertuna, Hacettepe University; İbrahim Uysal, Hacettepe University; Gunes Ertas, Bogazici University; Hülya Kelecioğlu, Hacettepe University

This paper presents a simulation study which aims to investigate the effects of differential item functioning (DIF) on the estimation of ability parameters by comparing different DIF determination methods: SIBTEST, IRT-LR, Lord’s χ,2, Raju’s area measure. DIF may cause the bias on the ability estimations (Camili, 1993), so even if item purification threatens the validity of the test, it may be a solution to more precise ability estimation (Golia, 2010). This study investigates the effects of DIF on ability estimation by purifying items that show DIF under different conditions. For this study three factors were examined: the rate of DIF (10 %, 20 %), the level of DIF (B&C and C) and the type of DIF (non-uniform, uniform, both non-uniform and uniform). For the purpose of study, initial item and ability parameters were derived for 1000 participants and their response to 30 dichotomous items with 50 replications. Considering different DIF determination methods, items showing DIF were detected, purified and then new abilities were estimated. The error and fit indices (RMSD and Pearson correlation) were calculated for old and new estimated abilities. The results show that while the rate of DIF increase the error decrease. The least error is for the method of IRT-LR on non-uniform 10% DIF condition. The highest correlation is also for the method of IRT-LR on both non-unifom and uniform 10% DIF condition. For SIBTEST and IRT-LR, only C level items' purification provides getting better estimation compared to purifying B and C level items under the all conditions.

DIF 2b: Impact of Decreasing Category Number of Polytomous Items on DIF

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

Although polytomous items consist of more detailed information and show more reliable results; these items can be recoded into less-categorized or even dichotomous scored items in practice. In a study conducted by Zeng, Walker, Tang and Potter (2015), polytomous items were recoded as dichotomous, DIF analyses were conducted, and compared, for both the original items and the recoded items. The present study analyzes the change in DIF amount of items in the case of a gradual decrease in the number of categories for the polytomous items. To this end, attitude towards science, from PISA 2006 were used. Attitude items consisted of polytomous, Likert-type statements. Polytomous items were re-coded as trichotomous and then as dichotomous. Afterwards, poly-SIBTEST and ordinal logistic regression method were used to test for gender DIF within the US sample. In this study, various coding methods were used and the results were assessed by standard errors. This study is of importance for revealing the effect of the decreasing the number of polytomous categories in DIF analyses using a large-scale real dataset.

DIF 2c: Assessing Measurement Invariance in TIMSS Using BSEM and Alignment Methods

Xinya Liang, University of Arkansas; Wenjuo Lo, University of Arkansas

Evaluating measurement invariance across diverse educational contexts is challenging in international studies. Two alternatives, Bayesian structural equation modeling (BSEM; Muthén & Asparouhov, 2013) and the alignment method (Asparouhov & Muthén, 2014), have drawn attention for multiple-group comparisons in confirmatory factor analysis. BSEM specifies informative priors on differences in measurement parameters, which allows approximate measurement invariance. Measurement non-invariance is detected, if parameter differences across groups are significant. These parameters may be freely estimated in subsequent analyses. The alignment method minimizes the amount of non-invariance through a simplicity function by considering differences in factor means and variances. Configural intercepts and loadings are estimated first, followed by the alignment procedure to obtain aligned parameter estimates and standard errors. Currently, little investigation has been done on empirical comparisons of the two methods. Our study purpose was to empirically investigate measurement invariance using BSEM and alignment methods.

This study used the TIMSS 2011 student value mathematics scale in eighth grade from North American and European countries. A one-factor model was used for the 6 items measured on a 4-point Likert scale. For the BSEM method, a prior distribution N(0 .05) was used for parameter differences. For the alignment method, Bayesian estimation was used with non-informative priors. Results were evaluated based on model fit, non-invariant parameters, and estimated factor means. This study is unique by making empirical comparisons of BSEM and alignment methods using international data. Results from the current comparison could benefit practitioners involved in international studies.

Diagnostic Classification Model- DCM 1: 4:25 PM - 5:55 PM

DCM 1a: Estimating Mixture Fit Index for Cognitive Diagnosis Models

Kevin Carl Santos, University of the Philippines; Jimmy de la Torre, The University of Hong Kong; Matthias von Davier, Educational Testing Service

Rudas et al. (1994) proposed the use of a mixture fit index as goodness-of-fit measure in contingency tables. It assumes that observations can be classified into two groups, namely: those that conform to a parametric model and those that do not. The mixture fit index gives the proportion of misfitting observations. This study applies the mixture fit index to detect aberrant response patterns in the cognitive diagnosis model (CDM) framework. Using nonlinear programming and bisection method, the proposed algorithm iteratively estimates the mixture fit index and posterior probabilities are then used to calculate the likelihood that response patterns belong in the aberrant group. Preliminary results show that the proposed procedure can identify aberrant response patterns for short tests.

DCM 1b: Hypothesis Testing for Item Consistency Index in Cognitive Diagnosis

Lihong Song, Jiangxi Normal University; Wenyi Wang, Jiangxi Normal University; Shuliang Ding, Jiangxi Normal University

In conjunctive condensation rule, hierarchy consistency index (HCI) or item consistency index (ICI) can be used to assess whether actual response patterns match the expected response patterns. However HCI or ICI cannot be directly used with disjunctive rule where the mastery of all the attributes measured by an item is not necessary for successful performance. This leads us to propose a new index that is specifically designed to identify misfits of item response vectors for disjunctive rule. In addition, HCI or ICI makes decisions based on an individual response. Such an inference is often unreasonable. Considering statistical inference is generally more precise than everyday inference, we introduce a consistency index based on hypothesis testing to help detect misfitting item response vectors. To investigate whether the new index can work well under certain conditions, four important factors were included in a simulation study with five independent attributes: (a)cognitive diagnostic models including two conjunctive models and a disjunctive model; (b)the quality of test Q-matrix with the error rates from 0.1 to .4 with step .1, (c)the quality of items, and (d)the number of examinees with N = 500 or 1,000. The results showed that these indices can be used to evaluate cognitive assumptions, to assess the quality of test Q-matrix, to identify poor items with attribute misspecification, and so on. The new indices with original indices may provide better understanding of nature of cognitive assumptions, and help determine which psychometric model is more appropriate for a diagnostic test.

DCM 1c: The Comparison of DINA and RDINA Models Under Changing Conditions

Ömür Kaya Kalkan, Hacettepe University; Hülya Kelecioğlu, Hacettepe University; Tahsin Oğuz Başokçu, Ege University, Turkey

The DINA model is a cognitive diagnosis model (CDM) that has received growing interest from researchers. Applying CDMs to the fraction subtraction data (Tatsuoka, 1990), revealed problems related to classification of examinees, latent class sizes, and the use of higher order models (DeCarlo, 2011). Additionally, selecting the most appropriate model assumes critical importance if there are several models available that are appropriate for the data. The RDINA model was obtained by the reparameterization of DINA as a latent class logistic regression model and the HORDINA model was obtained by adding the θ to the model by DeCarlo (2011). In the present study, DINA–RDINA and HODINA–HORDINA models were compared under changing conditions (i.e., number of attributes, g and s parameter values, and number of items). The results show that for conditions where the g–s parameters values and number of attributes were low (respectively 0.1 and 3), reparameterized models generated identical values to the DINA models. However, when the g–s parameters values and number of attributes were increased (respectively 0.5 and 5), parameter estimations obtained from the models, latent class sizes, and AIC and BIC information criteria show differences through the values from the models. For the AIC and BIC indexes obtained from simulated and real data sets, RDINA and HODINA provide smaller values in comparison to the DINA and HORDINA models, respectively. Consequently, considering all the conditions, it was found that the BIC information criteria provided more consistent results in comparison to the AIC.

Multilevel/Hierarchical/Mixed- MLM 1: 4:25 PM - 5:55 PM

MLM 1a: Bayesian Longitudinal Binary Models with Subject-Specific Inference

Kevin Duisters, Leiden University; Elise Dusseldorp, Leiden University; Stef van Buuren, TNO, Utrecht University; Aad van der Vaart, Leiden University; Jean-Paul Fox, University of Twente

Implying individual latent curves from longitudinal binary data may suffer from high noise levels when sample sizes per visit are low. Typical examples are longitudinal educational testing or health monitoring studies with an imbalanced, sparse design. A Bayesian Hierarchical Linear Model is suggested to parametrize a reference prior towards which subject-specific estimates can shrink in case of low information availability. The central idea is to model the relation between time and the latent variable as (higher than one degree) polynomial. The proposed Empirical Bayesian method retrieves the posterior (in a classically difficult problem due to the dimension of integration) without relying on subjective prior information. Based on mean posterior estimates the reference prior is established once, which can then be used for personalized inference on any new subject within seconds in a fully Bayesian paradigm. A normal mixture approximation to the logistic distribution is incorporated in a Gibbs framework that generalizes to both the logit and probit link. Formulating the latent level as Linear Mixed Model with indirect variance specification including a Gaussian Process allows for non-equally spaced observations on a continuous time domain. When covariates are present in the hierarchy, individual latent curves with posterior credible bands can be visualized against peers on a covariate corrected reference diagram. The methods are broadly applicable, but illustrated on a dataset of infant developmental indicators gathered in The Netherlands, where Item Response Theory (IRT) models are used to display individual ability curves.

MLM 1b: Relationship Between Spatial and Reasoning Ability: Gender as a Moderator

Liu Yue, Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University; Zhang Danhui, Beijing Normal University

Gender has been well recognized to play an important role in influencing students' spatial abilities. It was also discovered that boys seemed to have advantages over girls in spatial cognitive skills (Swarlis, 2008). However, an international report on mathematics ability of fourth grade students stated a contradictory conclusion which identified that boys did not always score highly than girls in the geometric field (TIMSS, 2011). Therefore, the effect of gender on students' spatial ability is worth further exploration. Studies have shown that spatial abilities are affected by the reasoning ability to a great extent, and girls tend to solve spatial problems using reasoning strategies (Xu Yan, & Zhang Houcan, 2000).The current study aims to clarify to what extent students' reasoning ability can predict their spatial ability and how gender plays a role as the moderator in this progress. Hierarchical linear model were used in the datasets which are from a national wide assessment of Chinese students' mathematics skills in 2014, along with the background factors. Compare to boys, reasoning ability may have more predictive effect on spatial ability for girls. Some suggestions are provided for differentiated instruction between genders' gap according to our findings, such as paying more attention to cultivate the reasoning strategy for boys, others strategies for girls.

MLM 1c: Nonparametric Mixed-Effects Models for Psychological Data

Nathaniel E. Helwig, University of Minnesota

Linear mixed-effects (LME) regression models are a popular approach for analyzing correlated psychological data. However, LME models require one to know the parametric form of the relationship between the predictor variables and the response variable (e.g., the relationship between X1 and Y is linear given the other predictors). In many psychological applications, the form of the relationship between the predictors and the response is unknown and must be estimated from the data. Nonparametric (NP) extensions of the LME model have been proposed (Gu & Ma, 2005; Wang, 1998; Zhang et al., 1998), but the heavy computational burden makes these extensions impractical for analyzing large samples of correlated data, which are typical in psychology and education. In this talk, we discuss how recent computational advances (Helwig, 2015; Helwig & Ma, 2015) make it practical to apply NP mixed-effects models to large samples of correlated data. We present simulation results revealing the benefit of the new method, as well as examples demonstrating the potential of NP mixed-effects regression for psychological data. Our talk reveals that data-driven methods such as NP mixed-effects regression can provide novel insight about psychological functioning without requiring a priori information about the nature of the underlying process.

MLM 1d: Mixed-Effect Models for Assessing Interrater Reliability and Its Moderators

Patricia Martinkova, The Czech Academy of Sciences; Dan Goldhaber, University of Washington-Bothell

Interrater reliability (IRR), commonly assessed by intraclass correlation coefficient, is an important statistic for describing the extent to which there is consistency amongst two or more raters in assigned measures. Traditionally, IRR is estimated in fully crossed or nested designs with the use of analysis of variance (ANOVA)-based methods. In teacher selection, the data structure is often hierarchical and designs deviate substantially from the research ideal of a fully crossed design. Previous research has also shown some evidence of existence of moderators of IRR in selection instruments, such as rater experience and training.

In presented work, we build mixed-effect models to estimate IRR in complex multilevel context and to test hypotheses about moderators of IRR. We compare available estimation techniques in a simulation study. Models are applied on a dataset of ratings of teacher applicants to a school district. We estimate within school and across school IRR and we test hypotheses about possible moderators of IRR, such as applicant type (internal or external), rater experience and job category. We also enumerate the direct effect of IRR on predictive power of the selection instrument and its components as measured by teacher value added, and we offer practical applications and policy implications of the methods we employ.

MLM 1e: Sample Size Determination for CRTs with Randomization at any Level

Satoshi Usami, University of Tsukuba

Hierarchical data (also called multilevel data or clustered data) is common in behavioral research when data about lower-level units (e.g., students, clients, repeated measures) are nested within clusters or higher units (e.g., classes, hospitals, individuals). In a typical longitudinal experimental design, randomization is not performed at the lowest level due to procedural restrictions or training effect. The present research derives closed-form sample size determination formulas that can be used to ensure the desired statistical power for cluster randomized trials (CRTs) with continuous outcomes. This allows randomization to be performed at any level. The formulas in this paper are derived under the random intercept model for four-level data, with both balanced and unbalanced designs considered. Additionally, we address several theoretical results about standard errors of the experimental effect estimate, including the relative impacts of intraclass correlations at each level and its generalized expression with any-level randomization under any number of levels.

Wednesday, July 13, 2016

Item Response Theory- IRT 3: 8:00 AM - 9:30 AM

IRT 3a: A Modified Binary Optional Randomized Response Technique (RRT) Model

Jeong Sep Sihm, The University of North Carolina at Greensboro; Sat N. Gupta, The University of North Carolina at Greensboro, NC, USA

We propose a modified two-stage binary optional randomized response technique (RRT) model in which respondents are given the option of answering a sensitive question directly without using randomization device, if they think they don't need extra measures to protect their privacy. This particular approach was proposed by Sihm & Gupta (2015) by using the split sample method. However, instead of using the split sample method to estimate two unknown parameters from a sample, we now propose in this work to ask two separate questions of the same sample. First with question 1, we estimate the level of the sensitivity of the main research question by using the indirect question RRT model of Warner (1965). Then, by using the model of Sihm & Gupta (2015), the prevalence of the sensitive characteristic which corresponds to the main research question is estimated from the same sample with question 2. Simulation results of this new model are compared with those from a more realistic version of the optional binary unrelated question model of Sihm, Chhabra, & Gupta (2016). Computer simulation shows the new two-stage binary optional RRT model in this study gives smaller variances of the estimator for the prevalence of the sensitive characteristic than the other model does, when they have the same sample size. In sum, the proposed model allows us to have a smaller sample size for the same amount of variance of the estimator than other models, thus increasing practical appeal for survey workers.

IRT 3b: Estimation of Ability with Reduced Asymptotic Mean Square Error

Haruhiko Ogasawara, Otaru University of Commerce

A method of the weighted score or penalized likelihood for estimation of ability reducing the asymptotic mean square error is derived. In this method, associated item parameters are assumed to be given or estimated by using a separate calibration sample with the size of an appropriate order. The method can be seen as an extension of the weighted likelihood method that removes the asymptotic bias of the maximum likelihood estimator. In the proposed method, some bias is retained while variance is reduced by using a multiplicative constant for the weight in the weighted score. A lower bound of the constant minimizing the asymptotic mean square error is found under the logistic model having identical items. The lower bound is numerically also shown to be reasonable in the case of the 3-parameter logistic model, with and without model misspecification.

IRT 3c: A Comparison of Two MCMC Algorithms for 2PL IRT Model

Meng-I Chang, Southern Illinois University-Carbondale; Yanyan Sheng, SIUC

Markov chain Monte Carlo (MCMC) techniques have become popular for estimating item response theory (IRT) models. The current development of MCMC focuses on two major algorithms: Gibbs sampling (Geman & Geman, 1984) and the No-U-Turn sampler (Hoffman & Gelman, 2014), which are implemented in two specialized software packages JAGS (Plummer, 2003) and Stan (Stan Development Team, 2016), respectively. In order to make informed decisions about which MCMC algorithm to use when it comes to the fully Bayesian estimation, it is desired that extensive studies are conducted to compare parameter estimates from these competing Bayesian algorithms. To date, however, no research has looked at their comparison with an IRT model. This study focused on the two-parameter logistic (2PL) model by comparing the performance between Gibbs sampling and No-U-Turn sampler on the accuracy of different prior specifications for the discrimination parameter aj. Specifically, three priors were considered for aj: (1) log normal, i.e., aj ~ lognormal (0, 0.5), (2) truncated normal, i.e., aj ~ N(0, 1)T(0,), and (3) via a transformation to αj, where aj is assumed to be exp (αj) and having a standard normal prior for αj. Results suggest that Gibbs sampling performed similarly to the No-U-Turn sampler under most of the conditions considered. In addition, both algorithms recovered model parameters with a similar precision except for small sample size situations. Findings from this study also shed light on the use of JAGS or Stan with more complicated IRT models.

Factor Analysis- FAC 2: 8:00 AM - 9:30 AM

FAC 2a: Assessment of Horn's Parallel Analysis with Missing Data

Francisco Abad, Universidad Autónoma de Madrid; María Nieto, Universidad Autónoma de Madrid; Luis Garrido, Universidad Iberoamericana; Vicente Ponsoda, Universidad Autónoma de Madrid; Miguel Sorrel, Universidad Autónoma de Madrid

Horn's Parallel Analysis (PA) has consistently shown to be a superior method for the assessment of latent dimensionality, a critical phase in the validation of measurements and the development of theory. Even though PA has been extensively evaluated across numerous data structures and variable characteristics, not much is known about its performance in the presence of missing values. Due to the saliency of missing values in many real-world research scenarios, it becomes relevant to understand the performance of PA in these conditions. Thus, the current study assessed the accuracy of PA in conjunction with three missing value treatment methods, Listwise Deletion (PA-LD), Pairwise Deletion (PA-PD), and Full Information Maximum Likelihood (PA-FIML), across a wide range of factor structures. Seven variables were manipulated using Monte Carlo methods: Sample size, number of factors, number of variables per factor, factor loadings, factor correlations, missing data mechanism, and percentage of missing values. The results indicated that PA-LD was not a suitable technique due the great number of discarded cases, while PA-PD and PA-FIML produced good levels of accuracy, provided the sample sizes were medium or large. Under the most adverse conditions (small samples combined with large percentage of missing values), however, PA-PD outperformed PA-FIML and there was a non-negligible loss in accuracy in comparison to PA with the full matrices of data. The performances of the PA variants are discussed in light of the theoretical properties of the missing data treatment methods and practical guidelines are offered.

FAC 2b: A Meta-Analysis of the Factor Structure of the CLASS

Hongli Li, Georgia State University; Jingxuan Liu, Georgia State University; Charles Vincent Hunter, Georgia State University

The Classroom Assessment Scoring System (CLASS) is a classroom observation instrument that has been extensively used to measure teacher-student interaction and classroom instructional quality. Based on the developmental model of learning, the CLASS has three primary domains. Domain One is Emotional Support (four dimensions); Domain Two is Classroom Organization (three dimensions); and Domain Three is Instructional Support (four dimensions). Researchers typically use the three domain scores of CLASS. However, previous validation studies did not find sufficient evidence to support this factor structure (i.e., each domain is one factor). Given the large number of studies using the CLASS in different settings and a pressing need for a more general understanding of the CLASS factor structure, we propose to conduct a meta-analysis of the CLASS factor structure. Following Cheung and Chan's approach, our meta-analysis involved two steps. The first step was to collect correlation matrices of the CLASS dimensions and then integrate them to a pooled matrix. Using the keyword "Classroom Assessment Scoring System," we searched well-known databases for articles using the CLASS and contacted the leading author if the identified article did not contain a correlation matrix. The second step was to examine the factor structure of CLASS using the pooled matrix. A few possible factor structures were investigated to examine the fit to the data. This meta-analysis is expected to provide more general empirical evidence to the CLASS factor structure. It has important implications to practitioners and researchers in the field of classroom instructional quality.

FAC 2c: Frequentist Model Averaging in Factor Score Regression

Shaobo Jin, Uppsala Univesrity

Traditional application of statistical modelling consists of several steps, one of which is model selection. The “best'' model is chosen from a class of candidate models. In factor score regression with exploratory factor analysis, factor scores are used as explanatory variables to model the observed response variable. The number of factors has to be decided according to a formal selection criteria. However, an incorrect factor number leads to severe consequences. In this study, we consider an alternative approach, namely, the model averaging approach, where all candidate numbers of factors are considered. Instead of relying on the estimated number of factors, models with all candidate factor numbers are fitted. We show that, if model averaging is introduced to the number of factors, the model error with properly chosen weights asymptotically achieves the infimum. Thus, if prediction is the primary purpose, model averaging over the factor number is an appealing alternative.

FAC 2d: Validity Analysis Using Factor Analyses on Defining Issues Test

Youn-Jeng Choi, University of Alabama; Meghan Bankhead, University of Alabama; Stephen Thoma, University of Alabama

The Defining Issues Test (DIT) is a test concerned with how one defines the moral issues in a social problem (Rest, Narvaez, & Thoma, 1999). Since Rest (1974) devised the DIT, and Rest, Narvaez, & Thoma (1999) revised it into the DIT-2, this test has been widely used as one of the most popular means of measuring moral judgment development. Although DIT researchers have developed extensive evidence to support the validity of the measure (Rest, Narvaez, Bebeau, & Thoma, 1999), much less is known about validity as it relates to the internal structure of the DIT. The proposed study will analyze validity through exploratory/confirmatory factor analysis. The second-order factor analysis and bi-factor analysis will also be applied using Mplus computer software. The DIT-2 datasets for this proposed study were already collected by The Center for the Study of Ethical Development, located at the University of Alabama. The samples were collected from undergraduate students in the United States during 2000 - 2009 (N = 49,611). Findings of the proposed study will provide further insight into interpretation of DIT scores for the evaluation of moral judgment. This validity study may provide a new and modified rubric to measure moral judgment development, and strengthen the results of previous validity studies.

Bayesian Statistical Inference- BSI 1: 8:00 AM - 9:30 AM

BSI 1a: Bayesian Inferences of a Rater Accuracy Unfolding Model

Fei Zhan, BPA Quality

Often in rater-mediated assessments, rater accuracy is checked by comparing observed and expert ratings. Recently, the hyperbolic cosine model (HCM) is proposed to unfold dichotomous accuracy ratings into different latent accuracy categories: inaccurate below expert ratings, accurate ratings, and inaccurate above expert ratings. Previous research suggests that HCM can provide a useful interpretive framework for assessing rating quality. However, the statistical treatment of the model is limited. So far, only joint maximum likelihood estimation of the model has been discussed in literature. In this paper, we introduce a Hamiltonian Monte Carlo (HMC) algorithm for fitting the model. The Hamiltonian Monte Carlo (HMC) algorithm simulates Hamiltonian dynamics by using a leap frog method. In general, it converges faster to the target distribution than most other MCMC samplers. And the algorithm is implemented in the general-purpose Bayesian package STAN. A simulation study is performed to examine the parameter recovery. In addition, posterior predictive model checks are used to evaluate model fit. At the end of the paper, we will also demonstrate the application of the model by using a real data set.

BSI 1b: A Bayesian Approach to Data Fusion

Katerina Marcoulides, Arizona State University; Kevin Grimm, Arizona State University

The exponential growth in the amount of data collected on individuals has created great opportunities for examining new theories and developing new methods of analysis. At the same time, the number of different sources over which this information is divided continues to grow, creating numerous obstacles for effectively combining such data before it can be explored. At its most basic level the process of combining data is one in which information from disjoint data sets sharing at least a number of common variables is merged. The process is commonly referred to as data fusion (Marcoulides & Grimm, 2015). Determining appropriate methods for pooling independent samples from different sources for simultaneous analysis continues to be problematic.

A variety of methods for data fusion have been proposed in the literature. To date, most are employed within the frequentist framework (e.g. Hofer & Piccinin, 2009). This study introduces a novel Bayesian approach for data fusion. The approach employs several separate but well-defined steps performed within a traditional Bayesian framework. However, instead of simply combining data sets at once, information obtained from one data set acts as priors for the next analysis. This process continues sequentially until a posterior distribution is obtained that incorporates information from each data set. Such an approach not only provides more accurate parameter estimates than other approaches typically used, but also leads to more accurate interpretations of obtained results. To illustrate its effectiveness, the method is applied to the fusion of simulated and real data.

BSI 1c: A Novel Method to Numerically Compute Parametrization Invariant Priors.

Merijn Mestdagh

There are at least two approaches within Bayesian statistics for constructing a prior. Subjective Bayesians try to express in the prior their beliefs about the parameter values based on information from past experiments. Objective Bayesians aim to construct priors with certain desirable features, such as requiring that the specific parameterization of the likelihood does not influence the inferred results. The best-known example of such a parametrization invariant prior is the Jeffreys prior. Unfortunately, for many interesting likelihoods the analytical derivation of most parameter invariant priors is difficult, if not impossible. In this presentation, we propose a new method to numerically compute a broad class of parametrization invariant priors, including but not limited to Jeffreys prior.

The presentation consists of three parts. In a first part, some existing but lesser known results concerning parametrization invariant priors will be explained. It is demonstrated that requiring all distinct distributions to be of equal importance, as assessed under some distance between likelihood distributions, systematically leads to a prior that does not depend on the particular parametrization of the model in question. Choosing the Kullback-Leibler divergence as the distance between a models likelihood distributions, for example, leads to Jeffreys prior. In a second part, these theoretical results are exploited to facilitate the calculation of a broad class of parametrization invariant priors, including Jeffreys prior. This results in a well-defined numerical procedure for deriving parameter invariant priors. Finally, the newly proposed procedure is evaluated in benchmark examples where analytical expressions for Jeffreys prior are available.

BSI 1d: Controlling for Multiplicity Using Bayesian Multilevel Models with Semi-Informed Priors

Michael Zweifel, University of Nebraska-Lincoln, Weldon Smith, University of Nebraska-Lincoln

When hypotheses about several means are tested simultaneously the type I error rate is inflated. This may result in researchers making incorrect inferences. Several procedures have been developed to control the type I error rate at alpha; however these procedures sacrifice a large amount of power to do so. This power loss is made worse when variance heterogeneity is present and a large number of hypotheses are tested. Multilevel models provide some control for type I error inflation by virtue of shifting sample mean estimates closer to the aggregated mean; however, this does not address the variance heterogeneity problem. By adopting a Bayesian approach, it is possible to assign unique prior distributions to each of the level two variances in order to account for variance heterogeneity. The level two variances are assigned an inverse gamma prior distribution, where the hyper parameters are determined using the sample variance estimates. This performance of this method was then compared to traditional multiple comparison procedures. Results indicate that this method was more powerful than traditional methods, particularly when the number of hypotheses being tested was large. However, it was not possible to determine whether this method maintained strong control of the type I error rate at alpha.

Structural Equation Modeling- SEM 2: 8:00 AM - 9:30 AM

SEM 2a: Comparison of GSCA and Maximum-Likelihood for Small Sample Invariance Assessment

W. Holmes Finch, Ball State University; Brian French, Washington State University; Maria E. Hernandez Finch, Ball State University

Invariance assessment is important for scale validation. Empirical evidence is needed to demonstrate that the scores on an instrument have essentially the same meaning for all individuals in the population, regardless of their subgroup membership (e.g., gender). This is critical in order to make accurate inferences about individuals or groups based on the test scores. Typically, invariance assessment involves multiple groups confirmatory factor analysis (MGCFA) in which factor models have increasingly restrictive equality constraints placed on the model parameters across groups. When such constraints substantially degrade model fit, the researcher concludes that invariance is not present. Further analyses can assist with the identification of specific model parameters that lack invariance.

In certain instances, researchers are faced with small samples for one or more of the groups. In such cases, the standard maximum-likelihood (ML) estimator may yield biased parameter estimates. This issue is magnified in the context of MGCFA invariance testing, where the total sample is divided into two or more groups, thereby potentially compromising parameter estimation within a group and deleteriously impacting invariance assessment.

As an alternative estimator, generalized structured component analysis (GSCA) may remedy invariance testing problems associated with small samples. Prior research has shown that GSCA can accurately recover latent variable model parameters in many instances, including with small samples. The goal of this simulation study is to compare performance of GSCA and ML for invariance testing with small samples, across a variety of conditions. In addition, invariance testing with GSCA is demonstrated using an extant dataset.

SEM 2b: Apply Structural Equation Modeling to Errors-in-Variables System Identification

Fan Wallentin, Uppsala University

Errors-in-variables (EIV) identification refers to the problem of consistently estimating linear dynamic systems whose output and input variables are affected by additive noise. Various solutions have been presented for identifying such systems. In this study, EIV identification using structural equation modeling (SEM) is considered. Two schemes for how EIV Single-Input Single-Output (SISO) systems can be formulated as SEMs are presented. The proposed formulations allow for quick implementation using standard SEM software. By simulation examples, it is shown that compared to existing procedures, here represented by the covariance matching (CM) approach, SEM-based estimation provide parameter estimates of similar quality.

SEM 2c: A Lasso Estimator for Structural Equation Models with Interaction Effects

Holger Brandt, University of Tuebingen; Nora Umbach, University of Tuebingen, Germany; Kevin A. Fischer, University of Frankfurt; Jenna M. Cambria, University of Arkansas; Augustin Kelava, University of Tuebingen

Psychological theories often include a variety of variables and complex multivariate nonlinear relationships. For example, theories developed to predict educational choices often include multiple correlated predictors, such as vocational interests, self-efficacy, and academic achievements. In addition to linear relationships, interactions between self-efficacy, vocational interests, and academic achievements have been shown to increase the explanatory power of the model (e.g., Patrick, Care, & Ainley, 2011). Although vocational interests include six correlated facets, researchers tend to analyze separate models for each of these facets and typically test only one or two of the proposed interaction effects to avoid multicollinearity. Currently, it is unclear whether this procedure is adequate. Furthermore, simulation studies have only investigated minimal structural equation models (SEM) with two predictor variables and single interaction effects. It is not clear how approaches designed to analyze latent interaction effects react to multicollinearity when many interaction effects are tested simultaneously. Here, we provide evidence that the standard sequential analysis of subsets of predictor variables to test (latent) interactions leads to inflated type I error rates that cannot be attenuated by correction procedures for multiple testing, such as Bonferroni corrections. In order to overcome this problem, procedures that efficiently analyze numerous latent interaction effects simultaneously are required. We introduce a lasso estimator for SEM with interaction effects that has an increased power to detect effects compared to traditional procedures (like LMS or product indicator approaches). In a simulation study, we show that this advantage becomes more apparent with increasing multicollinearity.

SEM 2d: Comparison of Frequentist and Bayesian Regularization in Structural Equation Modeling.

Ross Jacobucci, University of Southern California; Kevin Grimm, Arizona State University; John J. McArdle, University of Southern California

In the context of regression, the relationship between frequentist and Bayesian approaches to regularization has been long recognized (Park & Casella, 2008; Tibshirani, 1996). However, regularization has not explicitly been applied to structural equation modeling (SEM) until the recent implementation of regularized SEM (RegSEM; Jacobucci, Grimm, & McArdle, In Press). RegSEM allows for the application of lasso and ridge penalties on a variety of parameters within frequentist SEM. The Mplus implementation of Bayesian SEM (BSEM; Muthén & Asparouhov, 2011), using small variance normal distribution priors, can be seen as form of Bayesian regularization using ridge penalties. Thus, the purpose of this study was to demonstrate both the equivalence and distinction when performing regularization with Bayesian and frequentist estimation in SEM. The methods were compared on the Holzinger Swinefored (1939) dataset using the model explicated in Muthén & Asparouhov (2011). For frequentist estimation, ridge and lasso penalties were performed with RegSEM (in the R package regsem), whereas for Bayesian estimation, ridge was performed in Mplus, and lasso in JAGS. Across frequentist and Bayesian estimation, just as in the context of regression, ridge estimates were the same. On the other hand, with lasso penalties, only RegSEM pushed the estimates to zero. This finding agrees with Park and Casella (2008), where Laplace distribution priors operated as a form of hybrid between the ridge and lasso. This study was the first to show both the differences and similarities between methods for performing regularization in SEM, thus adding to the dearth of application of regularization in psychometrics.

Diagnostic Classification Model- DCM 2: 8:00 AM - 9:30 AM

DCM 2a: General Q-matrix Refinement in Cognitive Diagnosis: A Nonparametric Approach

Yan Sun, Rutgers, the State University of New Jersey; Yanhong Bian, Rutgers, the State University of New Jersey; Chia-Yi Chiu, Rutgers, the State University of New Jersey

In cognitive diagnosis, the Q-matrix (Tatsuoka, 1985) is a common component used to describe the item-attribute relationships for all analyses. It is usually pre-specified by content experts and this subjective procedure may cause misspecifications, which can negatively affect the estimation of parameters and in turn the classifications of examinees (Rupp & Templin, 2008). Some researchers have proposed methods to detect and correct misspecified entries in the Q-matrix; however, these existing methods are either applicable only to some simple models, such as the DINA or the DINO model (e.g., de la Torre, 2008; DeCarlo, 2012), or require a cut-off value which is difficult to specify for real data (de la Torre & Chiu; accepted).

The current study proposes a general Q-matrix refinement method that can be used when data conform to complex models such as the saturated general models without extra ad-hoc specification. The method follows the nonparametric framework of the Q-matrix refinement method proposed by Chiu (2013) that identifies the correct q-vectors in the Q-matrix by minimizing the residual sum of squares (RSS). However, it is innovative in that it incorporates the general nonparametric classification (GNPC, Chiu & Sun, 2015) method to classify data that guarantees good classifications when data conform to the saturated general models.

Simulation studies were conducted to examine the performance of the proposed method by varying factors including sample size, test length, number of attributes, proportion of misspecified entries, and attribute pattern structure.

DCM 2b: Nonparametric Calibration of Item-by-Attribute Matrix in Cognitive Diagnosis

Youn Seon Lim, University of Illinois at Urbana-Champaign; Fritz Drasgow, University of Illinois at Urbana-Champaign

A key component in cognitive diagnosis models is the so-called Q-matrix, which specifies the relationships between items and attributes, so that responses to each item can provide diagnostic information about mastery of the attributes. Most commonly, expert opinion is utilized to define the elements in the Q-matrix, perhaps followed by extensive statistical validation. In this study, we have proposed an accurate and automated statistical method for estimating the elements of Q-matrices for new items, given a set of items with a known Q-matrix under the conditions of conjunctive, disjunctive, and compensatory attribute relationships. The method requires no assumed parametric cognitive diagnosis model, and provides consistent estimates under a very general class of cognitive diagnosis models. Moreover, the method can easily be adapted to parametric modeling settings, and may be more efficient if the assumed model is correct. A final benefit is that it may be used one item at a time to validate Q-matrices determined through expert opinion.

DCM 2c: Q matrix Estimation in Bayesian Approach

Yinghan Chen, University of Illinois at Urbana-Champaign; Steven A. Culpepper, University of Illinois at Urbana-Champaign; Yuguo Chen, University of Illinois at Urbana-Champaign; Jeffrey Douglas, University of Illinois at Urbana-Champaign

Cognitive diagnosis models are partially ordered latent class models and are used to classify students into skill mastery profiles. The deterministic inputs, Noisy 'And'' Gate (DINA) model is a popular psychometric model for cognitive diagnosis. Application of the DINA model requires content expert knowledge of a Q-matrix, which maps the attributes /skills needed to master a collection of items. Misspecification of Q has been shown to yield biased diagnostic classifications. We introduce a Metropolis-Hastings sampling algorithm for estimating the DINA model Q-matrix. The developed algorithm builds upon prior research (Chen, Liu, Xu, & Ying, 2015) and ensures the estimated Q-matrix is identified. Monte Carlo evidence is presented to support the accuracy of parameter recovery. We apply the developed methodology to Tatsuoka's fraction-subtraction dataset.

DCM 2d: Irreplaceability of a Reachability Matrix

Shuliang Ding, Jiangxi Normal University; Wenyi Wang, Jiangxi Normal University; Fen Luo, associate professor; Jianhua Xiong, associate professor

In cognitive diagnosis, Q-matrix is an important concept . A reachability matrix R is a special Q-matrix which represents the direct or indirect relationship among the attributes. And it can present the cognitive model. R can be obtained from the adjacent matrix and vice sersa. Moreover, the reachability matrix R has two important properties: One is that any column in Q-matrix can be expressed by a linear combination of the columns of R with the combination coefficients being 1 or 0, this can be proved through the augment algorithm (Ding,et al. 2008; Yang et al. 2011). The other is that under the conditions of 0-1 scoring rubric and the noncompensatory among the attributes, if R is a sub-matrix of the test Q-matrix (this test Q-matrix is called as a necessary Q-matrix), then the ideal response patterns corresponding to any two different knowledge states are different. It is proved that these properties of R are irreplaceability, i.e., any other Q-matrix does not have one of these properties. Obviously, the necessary Q-matrix is different from the sufficient matrix proposed by Tatsuoka (1995,2009). A counterexample is provided to explain that the concept of necessary Q-matrix instead of the concept of sufficient Q-matrix can promote the construct validity.

DCM 2e: Three-Step Estimation of Cognitive Diagnosis Models with Covariates

Charles Iaconangelo, Rutgers, the State University of New Jersey; Jimmy de la Torre, The University of Hong Kong

Applied researchers often seek to relate covariates or external variables to latent group classification. The literature on latent class models offers two approaches, the one-step or three-step procedure. The former incorporates the covariates in the latent class model, estimating the parameters for the measurement model and the structural model simultaneously. The latter estimates the parameters of the measurement model using only the item responses, then proceeds to use the examinee classification as observed dependent variables in a multinomial logistic regression. The three-step approach leads to downward bias in the parameter estimates (Bolck, Croon, & Hagenaars, 2004).

In a 2010 paper, Vermunt proposed a new three-step maximum likelihood procedure that used the examinee classifications from the second step as the dependent variable in a multinomial logistic regression that incorporated the measurement error probabilities. These measurement error probabilities were calculated as the conditional probability of the estimated classification, given the true classification. Including this matrix of conditional probabilities in the objective function corrects the bias in the third-step estimates. An improvement to this correction procedure is proposed for the cognitive diagnosis framework that can be applied at the level of the attribute vector or the individual attributes.

A simulation study is designed to investigate the ability of the revised correction matrix to estimate the parameters of the structural model. Factors manipulated include sample size, test length, the number of attributes, and item quality. Preliminary results suggest that the proposed procedure returns incremental improvements over a variety of conditions.

Multivariate Analysis- MVA 1: 8:00 AM - 9:30 AM

MVA 1a: Regressing Multivariate Similarities onto Covariates: Test Statistic and Effect Sizes

Daniel McArtor, University of Notre Dame; Gitta Lubke, University of Notre Dame; VU University Amsterdam

Pairwise similarities between subjects' multivariate response profiles can be regressed onto covariates using Multivariate Distance Matrix Regression (MDMR). MDMR is similar to conducting multidimensional scaling on a distance matrix computed from a multivariate outcome and jointly regressing all resulting principal coordinate vectors onto the set of covariates. MDMR is not commonly used in psychological research, which is partially due to the fact that computationally intensive permutation tests are required to compute MDMR p-values, as well as the fact that there are no methods for evaluating the size of the overall or individual predictor effects. We establish the null distribution of the MDMR test statistic such that effects can be evaluated without the need for permutation tests, and also present a new effect size measure for individual outcome items. The distribution of the test statistic and the effect size measure are validated using simulated data, and the utility of MDMR for psychological research is showcased in an empirical analysis in which similarities between multivariate personality profiles are regressed onto age, sex, education, self-rated health, and their interactions. MDMR is a powerful method that affords the ability to detect and describe more complex multivariate relationships between the predictors and outcomes.

MVA 1b: Minimum-Risk Point-Estimation for the Standardized Mean Difference Using Sequential Estimation

Ken Kelley, University of Notre Dame; Bhargab Chattopadhyay, The University of Texas at Dallas

The standardized mean difference is a widely used measure of effect size. In this article, we develop a general theory for estimating the population standardized mean difference by minimizing both the mean square error of the estimator and the total sampling cost. Fixed sample size methods, when sample size is planned before the start of a study, cannot simultaneously minimize both the mean square error of the estimator and the total sampling cost. To overcome this limitation of the current state of affairs, this article develops a purely sequential sampling procedure, which provides an estimate of the sample size required to achieve a sufficiently accurate estimate with minimum expected sampling cost. Performance of the purely sequential procedure is examined via a simulation study to show our analytic developments are highly accurate. Additionally, we provide freely available functions in R to implement the algorithm of the purely sequential procedure.

MVA 1c: The Most Predictable Criterion with Fallible Data

Seock-Ho Kim, University of Georgia

Hotelling's (1935) canonical correlation is the Pearson product moment correlation between two weighted linear composites from two sets of variables. The two composites constitute a set of canonical variates, namely, a criterion variate and a predictor variate. Many statistical analyses in psychometrics deal with fallible data that contain measurement errors. A method of obtaining canonical correlations from the true-score covariance matrix is presented and contrasted with Meredith's (1964) for which the disattenuated canonical correlations are obtained from the correlation matrix of fallible data. Because nearly all linear statistical procedures including multiple regression, discriminant analysis, and many others can be viewed as special cases of canonical correlation analysis, the method potentially has wide-range applicability especially when data contain a subset of variables with measurement errors. Illustrations are presented with data from earlier journal articles.

MVA 1d: Structural Equation Modeling Approach to the Canonical Correlation Analysis

Zhenqiu Lu, UNIERSITY OF GEORGIA; Fei Gu, McGill University

Canonical correlation analysis (CCA) is a generalization of multiple correlation that examines the relationship between two sets of variables. Spectral decomposition can be applied and canonical correlations and canonical weights are obtained. Anderson (2003) also provided the asymptotic distribution of the canonical weights under normality assumption. However, CCA uses the unit-variance normalization so it provides the canonical correlations only without covariances of canonical variates. And then the two-set canonical variate (CV-2) model is proposed and it is mathematically equivalent to CCA. CV-2 uses the unit-length normalization on canonical weight vectors. It is very flexible because it can provide both canonical correlations and the covariances of canonical variates. The canonical correlations obtained from CV-2 by further using unit-variance normalization on canonical variates are the exactly same as those obtained from CCA directly. However, CV-2 is hard to implement for applied researchers because there is no existing software available and it requires writing their own computer programs.

In this article, we propose a structural equation modeling (SEM) approach to CCA via the CV-2 model. Mathematical forms are presented to show the equivalence among these models. The weight matrix is obtained as the inverse of the loading matrix and the variance or standard errors of weights are calculated through the delta method. Different popular SEM software such as Mplus, EQS, CALIS, OpenMx are demonstrated to illustrate the application, and the results are compared with those obtained from Anderson’s (2003) formula. Related issues are also discussed in the last section.

MVA 1e:Replacing Localized Focus Groups with the On-Line Kellian Mind Explorer

Joe Whitehurst, High Impact Technologies; Joel Gardi, Georgia Tech Research Institute; Pieter M. Kroonenberg, Leiden University & The Three-Mode Company

The Kellian Mind Explorer is an application for replacing traditional focus groups such as used by commercial firms and political organizations. It is based on George Kelly's Personal Construct Theory (http://www.pcp-net.de/info/about.html). The application has the form of a repertory grid as proposed by Kelly in The Psychology of Personal Constructs published in 1955.

Using the Kellian Mind Explorer, typical questions which are asked of focus groups are administered on-line and the results are presented graphically in a browser. Typically focus-group participants have to travel to meet with a skilled moderator and a small number of other participants at considerable costs to the organizers. In sharp contrast, the Kellian Mind Explorer enables conducting focus groups on-line and thus an unlimited number of persons can participate from anywhere, anytime without the interference of interpersonal relationships and moderator influences.

Computer-Based Testing- CBT 2: 8:00 AM - 9:30 AM

CBT 2a: Mode Comparability Studies for a High-Stakes Testing Program

Dongmei Li, ACT, Inc.; Qing Yi, ACT Inc.; Deborah Harris, ACT, Inc.

Mode comparability between paper and online versions of a test cannot be simply assumed, especially for high-stakes assessments. This paper presents the research designs, major findings, and practical issues from a series of special studies intended to ensure score comparability, including an online timing study and two mode comparability studies.

A randomly equivalent groups design was used for all these studies: Students were randomly assigned to take the test under different timing conditions in the online timing study and were randomly assigned to take the paper or online test in both mode studies. Timing recommendations resulted from the timing study were re-evaluated in the subsequent mode study, which resulted in a modification of the timing decisions. The updated timing decisions were then implemented in the second mode study.

Similar analyses were conducted to examine score equivalency and construct comparability between the paper and online versions of the tests in the two mode studies, which had different timings for the online tests. Score equivalency was examined in terms of the similarity of test score and item score distributions between the two modes. In addition, measurement precision was compared between modes, and the item latency information for the online test items was also examined. Construct equivalency was examined by comparing the dimensionality and factor loadings and by examining differential item functioning between paper and online scores. Student responses to survey questions were also analyzed.

Results from these studies can inform decisions when tests transition from paper to online.

CBT 2b: Bayesian Sequential Detection of Learning

Sam Ye, University of Illinois - Urbana Champaign

Computer-based assessments provide a self-paced platform where it may aid or even alternate in-class instruction to master multiple skills. However, unattended instruction may lack in guidance when item selection does not reflect one's progress nor promote learning. Under cognitive diagnosis models, our objective is to give an examinee a sequence of items that first focuses on learning until enough evidence leads to administering items that primarily focuses on detecting mastery for a subsequent advancement. This can be viewed as a sequential change-point detection problem, and we purport to minimize the delay between when the learning takes place and when the mastery is detected. The attribute vector of an examinee is expected to display a complete mastery of all attributes by the end of the assessment. To this end, a Bayesian approach of computing the probability of mastery and CUSUM-based statistic is used to administer item selection and further determine detection of mastery. Simulations are conducted to explore item administration rules that promotes learning with minimal delay for a fixed false detection error.

CBT 2c: Approaches to Achieving Comparability of Computerized Adaptive and Linear Testing
Tianli Li, ACT, Inc.; Jie Li, ACT, Inc.

Many testing programs are considering moving from linear to computerized adaptive testing (CAT) due to increasing demands of ongoing test administration and requirement of limited testing time. Also for a period of time, CAT may coexist with its linear version to increase accessibility. The purpose of this study is to investigate approaches and provide guidelines of achieving score comparability between CAT and its linear version in this transition process.

There are two major changes in the transition from linear testing to CAT. First, in CAT, each examinee may take different sets of items depending on their abilities. Second, in CAT, theta scoring method based on latent ability estimate under an IRT model is preferred instead of commonly used number correct (NC) scoring in linear testing.

Previous studies either suggested a lack of comparability between linear testing and CAT without identifying reasons or only dealt with one of the two major changes.

This study will examine (1) methods of transitioning from NC to theta scoring that preserve statistical characteristics of the existing score scale of linear testing and (2) CAT designs that may achieve comparability with linear testing, including a fixed length CAT using targeted information selection algorithm approximating information of a linear test and a variable length CAT with a stopping rule of reaching the same precision as the linear counterpart, with combination of different theta estimates. This simulation study aims at disentangling the differences caused by the two changes and focusing evaluations on psychometric properties of reported scores.

Keynote: Some Recent Developments in the Analysis of Data with Missing Values
Roderick Little, University of Michigan

Chair: Carolyn Anderson

Missing data are a common problem in psychometric research. Methods for handling this problem are briefly reviewed, including (a) pros and cons of different forms of likelihood inference, specifically maximum likelihood, Bayes and multiple imputation; (b) penalized spline of propensity models for robust estimation under the missing at random assumption, and comparisons with other doubly-robust approaches; and (c) subsample ignorable likelihood methods for regression with missing values of covariates. I'll also discuss two aspects of a recent National Research Council study on the treatment of missing data in clinical trials, namely how missing data impacts the choice of estimand, and sensitivity analysis for assessing departures from assumptions of the primary analysis.

Bayesian Item Response Theory Using Stata
Chuck Huber, StataCorp LP

Chair: Chuck Huber

Item response theory (IRT) is a popular tool for assessing the properties of individual items on tests and other instruments. Some IRT models such as three-parameter logistic (3PL) models with separate guessing parameters can be difficult or impossible to estimate using maximum likelihood. This talk will review the basic ideas of IRT, introduce the concepts and jargon of Bayesian statistics and Markov Chain Monte Carlo (MCMC) using the Metropolis-Hastings algorithm, and demonstrate how to fit 1PL-5PL Bayesian IRT models using Stata.

Item Response Theory-IRT 4: 10:55 AM - 12:25 PM

IRT 4a: A Comparative Analysis of Psychometric Unfolding Models

Giulio Flore, Leiden University

We present the results of a comparative analysis of simulation data of four unfolding Item Response Theory (IRT) models: Multiple Stochastic Unidimensional Unfolding (MUDFOLD); Generalized Hyperbolic Cosine Model (GHCM); PARellELLogram Analysis (PARELLA); and Generalized Graded Unfolding Model (GGUM).

We assess the sensitivity of methods implementing each model to Differential Item Functioning (DIF), Local Dependence (LD), non-standard distributional features of model parameters, and different items' starting values information. The simulations are based on a new unfolding IRT model, called Generalized Squared Distance Logistic Model (GSDLM).

The experiments show that bias and precision of item estimates are most affected by local dependencies and by the distribution of person locations on the latent trait. The bias and precision of individual latent trait value estimates are most affected by skewed distributions of personal latent traits and by the type of estimation method. MUDFOLD is shown to be sensitive to changes in the distributions of the item and person latent traits.

These results identify three areas of research for a more general application of unfolding IRT models: 1) how to yield precise estimates for item and person latent traits scores for both symmetrical and asymmetrical distributions of the person latent trait; 2) more effective ways to determine starting values for items compared to methods currently used or suggested; and 3) development of more efficient estimation algorithms that can evaluate large datasets and filter out incorrect item orderings as input to the method.

IRT 4b: Polytomous IRT Models via Categorical Item Factor Analysis and GLMM

Joseph Olsen, Brigham Young University

We utilize a basic framework for polytomous item response theory (IRT) models, based on the type of category comparisons (cumulative probability, adjacent category, continuation ratio, or baseline comparison) used for ordered and unordered categories, and the selection of an appropriate link function and linear predictor. Choosing the linear predictor involves 1) deciding whether to use a common discrimination parameter, item-specific discrimination parameters, or discrimination parameters that vary by category boundaries within items, and 2) whether or not to parameterize the estimated category boundary thresholds as item location (difficulty) and offset parameters, and if so, whether to equate the offsets across items. Adjacent category models include the basic and generalized versions of the Rating Scale and Partial Credit models. We extend the Generalized Partial Credit Model allowing the item-specific discrimination parameters to also vary by category boundaries. Counterpart cumulative probability models include the Graded Response Model (GRM) and various restricted and modified versions of the GRM. We also present a corresponding set of sequential IRT models based on continuation ratio logits that are described in the IRT literature, but not generally available in standard commercial or freely available IRT software. We demonstrate how to estimate the various models in Mplus as both Categorical Item Factor Analyses and Generalized Linear Mixed Models, and show how to derive the traditional IRT parameters from corresponding parameters directly estimated by Mplus, obtaining standard errors, significance tests, and confidence intervals.

IRT 4c: Modeling Performance Data: Polytomous Multilevel Testlet Model vs. Cross-Classified Model

Larissa Smith, NBOME National Center for Clinical Skills Testing; Wei Xu, university of florida

An important challenge in performance assessment involves properly modeling data with nested structures (e.g., test items nested within testlets or examinees nested within sampling sites, classrooms, or schools). Standard IRT models are not robust to model item or person dependence. To date, researchers have proposed several strategies to address local item and/or person dependence in hierarchical data (e.g., Beretvas & Walker, 2012; Jiao & Zhang, 2015; Xie, 2014; Jiao et al., 2012; Wang & Wilson, 2005). Most of these models, however, are intended to model dichotomously scored items rather than polytomously scored items. To address this research gap, we proposed two different psychometric models from a multilevel modeling perspective that are intended for polytomous items and compared the performance of these models using both simulated data and the real test data from the COMLEX (Comprehensive Osteopathic Medical Licensing Examination)-USA Level 2-PE, administered by the National Board of Osteopathic Medical Examiners (NBOME). Specifically, we modified a four-level polytomous testlet model based on Jiao and Zhang (2015) and extended a cross-classified model based on Xie (2014) that both accounted for the dual local dependence. For the real data study, two datasets were drawn from examinees' responses to COMLEX-USA Level 2-PE between July 2014 and April 2015. For the simulation study, three testlet variances (.5, 1, 2) and two sample sizes (1000 or 2000) were manipulated. WinBUGS was utilized to implement the statistical analysis. Results from this study might be helpful in understanding the validity of performance assessments.

IRT 4d: Detecting Differential Item Functioning in Polytomously Scored Testlet-Based Items
Wei Xu, university of florida; Larissa Smith, NBOME National Center for Clinical Skills Testing

Standardized testing has been widely utilized to assess test takers' ability. To ensure test fairness and construct validity of test items in standardized testing, DIF analysis has always been conducted. Indeed, DIF detection has been an important strand of research in psychometric literature for the past several decades. Researchers have investigated DIF within the framework of testlet response theory (TRT) models that account for the nested test structure so that the item parameter estimates and DIF detection are more accurate. Most IRT-based DIF detection models are primarily intended for dichotomously scored items rather than polytomously-scored items that are relatively common in assessment tests. Our study extends Beretvas & Walker's two-level multilevel measurement model (MMMT-2) to allow modeling of polytomous item responses' functioning and compare it with the three-level polytomous testlet response model. The performance of these models are evaluated to fit the simulated data. Three testlet variances (0.5, 1, 2), two sample size (1000 or 2000) and one DSF & DTLF magnitude (0.4) are manipulated. WinBUGS was utilized to conduct the statistical analysis. Our study advances current research by proposing an IRT-based model that addresses LID when identifying DIF among testlet-based polytomous item. This research can help test developers and policy makers to make better decisions regarding whether an item should be included or excluded in various testlet-based assessments.

IRT 4e: Measuring Physical Attraction with the Multidimensional Generalized Graded Unfolding Model
James S. Roberts, Georgia Institute of Technology; Matthew E. Barrett, Georgia Institute of Technology; David King, Georgia Institute of Technology

This paper illustrates a multidimensional version of the generalized graded unfolding model, known as the MGGUM, using a novel study on the psychology of physical attraction. The MGGUM is a noncompensatory multidimensional IRT model that jointly represents both persons and stimuli in a latent space. For this application, a digital artist created images of virtual female models for display on a computer screen. Models varied with respect to waist, hip and bust size along with overall body weight. The former three variables were crossed whereas body weight was nested to maintain realism, and this yielded 81 digital stimuli. Pairwise similarity judgments of stimuli were obtained from subjects using a matrix sampling design in which every 20 respondents constituted a virtual subject. Group-level multidimensional scaling (MDS) suggested that four dimensions were inherent in these similarity judgments, and that these dimensions corresponded well to the attributes that were varied. All subjects also rated the attractiveness of each digital model using a 4-point response scale. These data will be analyzed with a 4-dimensional MGGUM via a Markov chain Monte Carlo estimation method. The results from the MGGUM representation will be described from an IRT perspective. Moreover, the stimulus locations derived from the MGGUM will be compared to those obtained from MDS as well as the true characteristics of the virtual models. Thus these comparisons will provide both (multi-method) psychometric and psychophysical interpretations.

Causal Inference and Mediation-CAU 1: 10:55 AM - 12:25 PM

CAU 1a: Causal Inference with Observational Multilevel Data: Investigating Heterogeneous Treatment Effects

Jee-Seon Kim, University of Wisconsin-Madison; Wen-Chiang Lim, University of Wisconsin-Madison; Peter Steiner, University of Wisconsin-Madison

Causal inference with observational data is challenging, as the assignment to treatment is often not random, and people may have different reasons to receive or to be assigned to the treatment. The multilevel structure adds complexity to the issue, where the assignment to treatment can be determined by various sources; for example, different district policies, school resources, teachers' evaluations, parents' and/or students' choices in educational settings. In multilevel analysis, therefore, it is critical to account for not only the nested structure of the data but also potential heterogeneity in selection processes and/or treatment effects. This study presents methodology for classifying level-one and level-two units into homogeneous "latent classes" with regard to class-specific selection and outcome models. The classification into homogeneous groups can take place at a cluster level and also at the lowest level, depending on the main sources of heterogeneity in the selection and outcome mechanisms. The talk introduces the methods, discusses their properties, and provides recommendations for the proper use of the techniques in practice.

CAU 1b: Moderation of Nonlinear Bivariate Relationships

Johnson C. Li, University of Manitoba

A common research question is the extent to which a linear relationship between two continuous variables (X and Y; e.g., intelligence and achievement) is moderated by a third dichotomized variable Z (e.g., gender). For example, finding the slope that regresses achievement on intelligence (or correlation between Y and X) differs for boys and girls has important implications for theory development (Smithson, 2012). Despite its popularity, conventional moderation analysis focuses on detecting linear moderation (Hayes, 2015), meaning that researchers may detect a significant result if, and only if, a slope difference (Δb) or a correlation difference (Δr) exists between the two groups. In practice, there are data scenarios that conventional moderation analysis cannot capture. If the best fitted line for the intelligence-achievement relationship is horizontal for boys, and there is a quadratic relationship for girls, then both Δb and Δr become zero and conventional moderation analysis may not reveal a moderation effect. I propose an innovative approach that utilizes curvilinear parameters to capture different shapes of relationships across each level of variable Z. In this study, I have tested my proposed method by defining and simulating 8 common nonlinear-moderation models and running simulations based on multiple sample sizes and sample size ratios. The results reveal that the proposed method better captures the shape and magnitude of nonlinear moderation as compared to the conventional method. I offer this new moderation analysis method as a trustworthy tool for detecting nonlinear moderation and discuss implications of its use in psychological research.

CAU 1c: Scoring Options for Assignment Variables in Regression Discontinuity Designs

Monica Morell, University of Maryland, College Park; Ji Seung Yang, University of Maryland, College Park

Regression discontinuity (RD) analysis (Thistlethwaite and Campbell, 1960) is used to estimate local average treatment effects for quasi-experimental designs when assignment to treatment is determined by location on an observed continuous variable. However, the "observed" variable is not truly observed if it is measured by observed indicators such as categorical item responses (e.g., pre-test scores, achievement level, social-economic status), then it contains measurement error that can impact treatment effect estimates. While the effect of measurement error in predictors in regression analysis is well known (e.g., Spearman, 1904), the effect of measurement error in assignment variable on the RD treatment effect is less explored. Empirical studies in econometrics (Battistin &Rettore, 2009; Hullegie, 2010) have concluded RD estimates will be unbiased if measurement error in the assignment variable is independent from treatment. In contrast, Schumaker (1992) reported measurement error in the assignment variable (within the classical test theory

framework) results in biased RD treatment estimates. The purpose of this study is to explore the impact of measurement error in an assignment variable on RD treatment estimates via two two-stage estimation methods that use summed or Expected A Posteriori (EAP) scores from item response models and a one-stage estimation using the Bayesian approach. A simulation study is conducted under different sample sizes, measurement conditions and magnitudes of treatment effect. Preliminary results indicate using summed scores from a long test can yield better coverage rates across different band widths while using EAP scores can be particularly beneficial when narrow bands are used.

CAU 1d: Modeling Multifaceted Constructs in Statistical Mediation Analysis

Oscar Gonzalez Jr, Arizona State University; David P. MacKinnon, Arizona State University

When evaluating a health intervention with statistical mediation analysis, researchers need to identify the important aspects of the mediating construct in the causal process. By distilling the mediating variable, researchers could save resources by not having to implement a complete program or a whole questionnaire battery when only part of it was previously successful in changing the outcome. This problem is relevant when the hypothesized mediator consists of multiple related facets; the broad definition of the construct and its more specific facets might relate differently to the outcome, which could undermine accurate conclusions in statistical mediation. Here we present simulation results on the statistical properties of the mediated effect when one of the construct's facets is the true mediator (characterized by the bifactor model), but the mediator is misspecified with conventional measurement models. The models investigated include those that only model the broad construct, only model the specific facet, or structures that model both of them simultaneously. Of particular interest is identifying the conditions under which the mediated effect would be detected and when it would have the least bias. Future extensions of the method and limitations will also be discussed. This study contributes to the largely unexplored area of measurement issues in statistical mediation analysis.

CAU 1e: When Randomization Fails: Adjusting for Baseline Group Differences

Patrick E. Shrout, New York University

On the average, randomization guarantees that experimental manipulations can be interpreted as causal effects because the treatment and control group are equivalent with regard to possible confounders. However, when studies have small sample sizes, as is common in social neuroscience, there is an increased chance that treatment and control groups will differ in important ways at baseline. When chance imbalance occurs at baseline, adjustment can be made in the analysis through ANCOVA or by analyzing difference scores. However, the investigator might find that opposite conclusions are obtained if the baseline differences are adjusted using difference scores (Post-treatment minus baseline) versus regressed change scores (ANCOVA with baseline as covariate). Using tools from modern causal analysis I show why these approaches can give different answers and I provide a clear recommendation for which approach should be avoided.

Diagnostic Classification Model-DCM 3: 10:55 AM - 12:25 PM

DCM 3a: Bayesian Estimation of Generalized NIDA Model with Gibbs Sampling

Aaron Hudson, University of Illinois at Urbana-Champaign; Steven A. Culpepper, University of Illinois at Urbana-Champaign; Jeffrey Douglas, University of Illinois at Urbana-Champaign

We develop a Bayesian formulation for a generalized version of the noisy inputs, deterministic, "and" gate (NIDA) model. In contrast to the restrictive NIDA model, which estimates 2K guessing and slipping skill parameters, the developed model estimates Kj+1 parameters for each item, where Kj is number of skills required to answer item j without guessing. The Bayesian model formulation employs a data augmentation strategy for estimating parameters using Gibbs sampling. Specifically, the response of individual i on item j, Yij, is modeled as a product of latent attribute responses, Xijk. Monte Carlo evidence supports accurate parameter recovery and additional results compare the developed model with the reduced RUM, DINA, and

NIDA models in terms of model fit. Applications are presented using Tatsuoka's fraction-subtraction dataset and responses to a mental rotation test.

DCM 3b: The Reduced RUM as a Logit Model: Parameterization and Constraints
Hans Friedrich Koehn, University of Illinois at Urbana-Champaign

The Reduced Reparameterized Unified Model (Reduced RUM) has received considerable attention among psychometric researchers studying diagnostic classificaiton models (DCMs) for educational assessment. Markov chain Monte Carlo (MCMC) or marginal maximum likelihood estimation relying on the expectation maximization algorithm (MMLE-EM) are typically used for estimating the Reduced RUM. Commercial implementations of the EM algorithm are available in the latent class analysis (LCA) routines of Latent GOLD and Mplus, for example. Fitting the Reduced RUM with an LCA routine requires that it be re-parameterized as a logit model, with constraints imposed on the parameters. For models involving two attributes, these have been worked out. However, for models involving more than two attributes, the parameterization and the constraints are nontrivial and currently unknown. In this study, the general parameterization of the Reduced RUM as a logit model involving any number of attributes and the associated parameter constraints are derived.

DCM 3c: On the Estimation of Standard Errors in cognitive diagnosis models
Michel Philipp, University of Zurich; Carolin Strobl, University of Zurich; Jimmy de la Torre, The University of Hong Kong; Achim Zeileis, University of Innsbruck

Cognitive diagnosis models (CDMs) are an increasingly popular method for assessing mastery or nonmastery of a set of fine-grained abilities in educational or psychological assessments. To statistically compare different versions of CDMs and to check model assumptions, several techniques are available that require a precise estimation of the standard errors (or the entire covariance matrix) of the model parameters. It is shown that the currently used calculation leads to underestimated standard errors because it only includes the parameters for the item responses, but omits the parameters for the ability distribution. We show that including those parameters in the computation of the covariance matrix consistently improves the quality of the standard errors. The practical importance of this finding is discussed and illustrated using real data examples.

DCM 3d: Joint Maximum Likelihood Estimation of Cognitive Diagnosis Models
Youn Seon Lim, University of Illinois at Urbana-Champaign; Fritz Drasgow, University of Illinois at Urbana-Champaign

In this study, we proposed a simulation-based method for computing joint maximum likelihood estimates of cognitive diagnosis model parameters, carried out by means of a combination of the simulated annealing algorithm and stochastic simulation of the hidden Markov chain. The central theme of the approach is to reduce the complexity of models to focus on their most critical elements. 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 removed by integration with this approach; the joint distribution of the latent attributes does not have to be specified, which reduces the number of parameters in the model. The Markov Chain Monte Carlo algorithm is used to simultaneously evaluate and optimize the likelihood function. This streamlined approach performed 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.

Classification, Clustering and Latent Class Analysis-CCC 1: 10:55 AM - 12:25 PM

CCC 1a: GDCM-MC: the Better Measurement Tool for Multiple-Choice Items

Yanyan Fu, The University of North Carolina at Greensboro; Oksana Naumenko, The University of North Carolina at Greensboro; Robert Henson, The University of North Carolina at Greensboro; Louis DiBello, University of Illinois at ChicagoJohn Sessoms, The Un

Diagnostic Classification Models (DCMs) are used for modeling the relationship between dichotomous multidimensional attributes and a set of item responses. The Generalized Diagnostic Classification Models-Multiple Choice (GDCM-MC DiBello, Henson & Stout, 2015) extend traditional DCM frameworks to polytomous items. Unlike the traditional item-based Q-matrix used for dichotomous DCMs, the GDCM-MC applies Q-matrices that specify the skills (or misconceptions) measured by each option of a given item. Therefore, more information can be extracted from a single item for GDCM-MC than from dichotomous models. As a result, it is possible that the GDCM-MC can yield better latent attributes classification accuracy than the analogous dichotomous DCM, given similar conditions.

In this study, authors will compare the GDCM-MC with the extended RUM (ERUM) link (DiBello et al., 2015) to the dichotomous Reduced RUM (R-RUM; Hartz, 2002) by means of simulation and a real data application. First, the similarities of the two models will be compared using a simulation study. The data simulated from the R-RUM will fit to the ERUM, and the data simulated from the ERUM will fit to the R-RUM. The correct classification rates (CCRs), which tells the differences between the true and estimated attributes, will be used to detect the differences of the two models. Second, both models with two different Q-matrices will be used with a multiple choice dataset that measures the skills and misconceptions of the examinees. After estimation, the CCRs and Kullback-Leibler Cognitive Discrimination Indices (Henson & Douglas,2005) will be compared and evaluated.

CCC 1b: The Wald Test for Empirical Q-Matrix Validation

Jimmy de la Torre, The University of Hong Kong

The Q-matrix that specifies required attributes for each item is a crucial component of cognitive diagnostic assessments (CDAs). However, conventional Q-matrix development involves some chance of subjectivity, which can result in validity concerns. The validity inferences from CDAs can be improved by statistical analyses. This study proposes a new procedure, Wald-Q, which is the adaptation of a statistical test – the Wald test. The Wald-Q is a multivariate hypothesis testing that simultaneously compares all possible q-vectors at the item level. A simulation study is carried out under varying conditions (i.e., sample sizes, test lengths, number of random misspecifications, and item qualities) to examine the viability of the Wald-Q. In addition, situations in which the true underlying restricted model is known and unknown are considered. Findings are reported as the proportions correct specifications retained and misspecifications corrected in the Q-matrix. Based on the preliminary results, when there are no misspecifications, the Wald-Q perfectly retains all the correct specifications for the item. When misspecifications are present, the retention rates for correct specifications are generally perfect with some exceptions under low quality items, short test lengths and small sample sizes; in addition, the correction rate for misspecifications is excellent, except when the item quality is low. Finally, the procedure is compared to an existing Q-matrix validation procedure, and is shown to provide better results.

CCC 1c: A Mixture Partial Credit Model with a Language-Based Covariate

Seohyun Kim, University of Georgia; Allan Cohen, The University of Georgia

A mixture partial credit model (MixPCM) can be used to classify examines into a number of discrete latent classes based on their performance on items scored in multiple ordered categories. The MixPCM also provides estimates of examinee ability within latent classes. Characterizing the latent classes, however, is not always straightforward, particularly when analyzing text from constructed responses. Latent Dirichlet allocation (LDA) is a statistical model that has been used to detect latent profiles in textual data. The profiles can be used to characterize documents, such as answers on a constructed response test, as mixtures of a small number of topics. This simulation study investigates how these profiles can be used as covariates for predicting latent classes in a MixPCM. The simulation study is presented to examine the impact of practical testing conditions, such as sample size, number of topics, and number of words, on the LDA parameter estimates and by the MixPCM with a covariate composed of profiles detected by the LDA.

CCC 1d: A Q-Matrix Validation for a Polytomous Cognitive Diagnosis Model

Wenchao Ma, Rutgers, The State University of New Jersey; Jimmy de la Torre, The University of Hong Kong

Cognitive diagnosis models (CDMs) have received increasing attention recently. A central component for most CDMs is the Q-matrix (Tatsuoka, 1983), which specifies the association between items and attributes of interest. The development of the Q-matrix is typically based on expert judgment; however, this process tends to be subjective and therefore, is prone to errors. Studies have found that the misspecifications in the Q-matrix may lead to inaccurate attribute classifications. To address this issue, a number of Q-matrix validation approaches aiming to detect and correct the misspecifications in the Q-matrix have been developed in the literature. Nevertheless, until now, there is no Q-matrix validation method available for the CDMs developed for polytomously scored items.

Recently, a sequential generalized deterministic inputs, noisy and gate model (sequential GDINA; Ma, de la Torre, & Sun, 2015) has been developed for graded responses. It assumes that item categories are attained in a sequential manner, and that attributes are associated with item categories explicitly. In this study, a Q-matrix validation approach using the Wald test in a forward manner was proposed for this model, to detect and correct the misspecifications in the attribute and category association. Various conditions including sample sizes, item quality, and the proportion of misspecification were controlled in this study. The false-negative and false-positive rates were examined. Preliminary results showed that the proposed Q-matrix validation method had low false negative rates. The false positive rates were also low even when 10% elements in the Q-matrix were mis-specified.

CCC 1e: Q-Matrix Misspecification and Item-Fit Assessment in Option-Based DCMs

Oksana Naumenko, The University of North Carolina at Greensboro; Yanyan Fu, The University of North Carolina at Greensboro; Bob Henson, The University of North Carolina at Greensboro; Lou DiBello, University of Illinois at Chicago; William Stout, Universi

Diagnostic Classification Models (DCMs) are constrained latent class models that reveal fine grained information about examinee proficiency. Most DCM applications to date use information from one multiple choice (MC) response option per item (i.e., the correct option) to determine mastery skill profiles for a sample of examinees. A new family of models termed the Generalized Diagnostic Classification Models for Multiple Choice Option-Based Scoring (GDCM-MC; DiBello, Henson, & Stout, 2015) allows the user to extract cognitive processing information from all MC options, allowing inferences about skills as well as misconceptions that examinees may possess.

In this study, we consider the effect of sample size, dimensionality, and two types of Q-matrix misspecification on the performance of fit index Di,h   (DiBello, Henson, & Stout, 2015) and newly proposed D’i,h . The two D indices were designed for use with the GDCM-MC, and represent the average discrepancy between the conditional model-based probability of an option and the corresponding empirical conditional probability. Additional dependent variables will be used to provide context: mean absolute difference (MAD; Henson, Roussos, Douglas, & He, 2008) fit index, parameter recovery via correlations between simulated and estimated data, as well as correct (mastery profile) classification rates (CCRs). The two types of Q-matrix misspecification considered are over/underfitting (e.g., requiring an attribute’s mastery when the item does not measure it) and attribute dependency (e.g., always requiring a set of attributes in tandem). Data generation and parameter estimation will be performed using the extended RUM-MC model (DiBello et al., 2015).

Multidimensional Item Response Theory-MIRT 1: 10:55 AM - 12:25 PM

MIRT 1a: Alternative Dimensionality Structures and Their Effect on Test Information

Brian Habing, University of South Carolina; Louis Roussos, Measured Progress

It is often desired to report a single score for educational and psychological assessments, even though many of them are (often intentionally) multi-dimensional. The implications of this lack of model fit have been widely discussed, and a classic result is that the amount of information about the target unidimensional latent trait is over-estimated (Thissen, Steinberg, & Mooney, 1989). One solution to this difficulty is the use of a testlet model (Wang, Bradlow, and Wainer, 2002). In these models, the test is treated as a compensatory multidimensional item response theory model, where theta denotes the target dimension, and the nuisance dimensions are treated as independent random effects. For example, reading and the contents of various paragraphs that are prompts to several questions each. This set-up conflicts with the concept of the latent trait best measured by an exam (e.g., Zhang & Stout, 1999; Habing & Roussos, 2003), and the two traits are necessarily different.

This paper examines three distinct testlet structures: one where the classic testlet-based ability parameter seems appropriate, one where the Zhang and Stout definition seems more appropriate, and one that is in between. Using direct calculation and simulation, we evaluate a variety of factors (such as dimensional correlations) to examine the differences in the three ability models and their information functions when applied to the same test. This includes cases where each of the three settings is used to define the underlying structure of the test.

MIRT 1b: Projective Item Response Theory: Overview and Application to PISA Data

Edward Ip, Wake Forest School of Medicine; Shyh-Huei Chen, Wake Forest School of Medicine; Yanyan Fu, The University of North Carolina at Greensboro; Tyler Strachan, University of North Carolina Greensboro; Terry Ackerman, The University of North Carolina

Test developers have a love-hate relationship with multidimensionality. On the one hand, the presence of rich and possibly multidimensional test contents could enhance the validity and appeal of a test. On the other hand, multidimensionality creates technical problems for assessment, particularly in the use of unidimensional models. The projective item response theory (IRT) is designed to extract measures for an intended construct of interest—e.g., math abilities—in the presence of multidimensionality in the test responses. The underlying idea of the projective scheme is based on the “integrating out” of nuisance dimensions by first fitting a multidimensional IRT (MIRT) model to the data. The result is a unidimensional model that allows functions such as test comparisons that contain different nuisance dimensions—e.g., two math tests of which one contains a verbal component and another a spatial component. The presentation will discuss both the strategic and technical aspects of the approach—that is, the advantages and limitations of the projective approach as compared to other scaling methods and the technical implementation of the procedure. Furthermore, critical issues such as sensitivity of the projected model to the specification of the MIRT model will be briefly discussed with some preliminary results. An application of the projective IRT to PISA data will be presented. Finally, further directions for investigating the projective IRT methodology, which is currently supported by the Institute of Education Sciences, will be discussed.

MIRT 1c: A General Bayesian Multilevel Multidimensional IRT Model for Dual Dependence

Ken A. Fujimoto, Loyola University Chicago

Item response theory (IRT) models for dual dependence simultaneously account for local item dependence (LID) and local person dependence (LPD). These models account for the LID portion of dual dependence by specifying general and nuisance dimensions in the latent trait space, with the dimensions assumed to be orthogonal. Such assumption addresses the identification issues that arises when attempting to estimate the correlations among these types of dimensions (Rijmen, 2009; Sheng & Wikle, 2009). Unfortunately, just because these correlations are not identified from an estimation perspective does not mean that empirical data sets support such structure (Fujimoto & Hwang, 2014; Jennrich & Bentler, 2012; Sheng & Wikle, 2009). When the orthogonal assumption is violated, the validity in interpreting the ability estimates corresponding to the general dimension(s) could be weakened.

For this presentation, I introduce a more flexible Bayesian multilevel multidimensional IRT (BMMIRT) model for dual dependence. With this model, the correlations among the general and nuisance dimensions can be estimated, thus providing a test of the orthogonal assumption, and it does so while simultaneously controlling for LPD. This model contains parameters and assigns prior distributions not seen in other models for dual dependence, thus overcoming the correlation identification issue. A simulation study indicates that, when the data are simulated under a nonorthogonal condition, the BMMIRT model displays superior predictive performance over other IRT models for dual dependence, with its estimated correlation matrix matching the generating matrix. The utility of this model is also demonstrated on an empirical data set.

MIRT 1d: Bayesian Estimation of a Multidimensional Four Parameter Model

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

High-stakes decisions are often made with low-incentive achievement tests that bear minimal implications for test-takers. In the absence of clear incentives achievement tests may be subject to the effect of careless errors, which reduces score precision for higher latent scores and distorts test-developers understandings of item and test information. A multidimensional four parameter normal-ogive (4PNO) model was developed for large-scale assessments and applied to dichotomous items of the 2011 National Assessment of Educational Progress (NAEP) 8th grade mathematics and reading tests. The results provide evidence that the probability of slipping exceeded five percent for 46.3% and 51.1% of the dichotomous mathematics and reading items, respectively. Furthermore, allowing for slipping resulted in larger item discrimination parameters, increased information in the lower-to-middle range of the latent trait, and decreased precision for scores one standard deviation above the mean. The results provide evidence that slipping is a factor that should be considered during test development and construction to ensure adequate measurement across the latent continuum.

MIRT 1e: Properties of Conditional Multinomial Models as Multidimensional Item Response Models

Carolyn J. Anderson, University of Illinois at Urbana-Champaign; Hsiu-Ting Yu, National Chengchi University

Conditional multinomial models have been proposed as latent variable models (Andersen, 1995; Anderson & Vermunt, 2000; Anderson, Verkuilen & Peyton, 2010; Anderson, 2013; Goodman, 1979, 1981; Hessen, 2012; Holland, 1990; Li, 2010); however, the theoretical underpinnings of the conditional models differ from those of standard latent variable models, including multidimensional item response theory (MIRT) models. Anderson and Yu (2007) studied conditional multinomial models as uni-dimensional item response models for dichotomous items. We extend Anderson & Yu (2007) to the case of multidimensional models for polytomous items. The properties of conditional multidimensional item response theory (conMIRT) models are studied theoretically and empirically. Even though there are theoretical differences, empirically the conMIRT models behave very much like traditional MIRTs, and in some cases better than traditional MIRT models. ConMIRT models are log-multiplicative association models and special cases of log-linear models, which are not necessarily downward collapsible as are traditional MIRT models. This is one property that will be theoretically and empirically discussed. Since conMIRT models do not require an assumption for the marginal distribution of the latent variables, empirical studies will be presented that compare conMIRT and traditional MIRT models under different marginal distributions. An example is presented where item responses are highly skewed.

Model Fit, Comparison and Diagnostics-FCM 1: 10:55 AM - 12:25 PM

FCM 1a: Goodness-of-Fit tests for Multivariate Categorical Data Using Complex Survey Data

Irini Moustaki, Department of Statistics, London School of Economics; Chris Skinner, London School of Economics

Data on multiple categorical variables are often collected in surveys in order to measure a smaller number of underlying dimensions. Maximum likelihood type estimation methods can be computationally burdensome because of the need to compute multi-dimensional integrals and we shall consider composite likelihood methods which reduce this computation and can be adapted to allow for weights and other features of complex survey sampling schemes. We focus on the problem of testing the goodness-of-fit of the model or of testing associated nested hypotheses. Because the contingency table can be sparse when the number of variables is large we consider methods which focus on lower order margins of the table. Such methods are sometimes called limited information tests. We consider their extension to complex survey data, including the Rao-Scott tests.

FCM 1b: Tukey-Hann Approach for Model-Data Fit of Persons in Rasch Measurement

Jeremy Kyle Jennings, University of Georgia; George Engelhard, University of Georgia

This study describes an approach for examining model-data fit for the dichotomous Rasch model using Tukey-Hann Person Response Functions (TH-PRFs). A method for smoothing functions was proposed by Tukey (1977), and it can be used to examine person response functions (PRFs). Douglas & Cohen (2001) proposed a root integrated squared error (RISE) statistic to examine model-data fit by comparing nonparametric and parametric PRFs. In this case, the TH-PRFs can be viewed as nonparametric PRFs, while the Rasch model defines the parametric PRFs. Data from a large university introductory statistics test (25 items) are used to demonstrate this approach (n=1,255). This study also compares the residual based infit and outfit with the RISE statistic. Preliminary results suggest that the RISE statistic and TH-PRFs provide a useful analytical and graphical approach for evaluating person misfit.

FCM 1c: Candecomp-Parafac Algorithms for Compositional Data

Michele Gallo, University of Napoli L'Orientale; Viola Simonacci, University of Napoli L'Orientale

Compositional data consist of vectors of positive values summing to a unit, or in general, to some fixed constant. In many applications these data are organized in three-way arrays and modeled using the Candecomp/Parafac method. The standard procedure for fitting the model is the Parafac-ALS, thanks to its attractive properties (bounded loss function, least-square solution, stable results). These merits, however, go together with different shortcomings. Specifically, the Parafac-ALS computation poses difficulties due to a non-convex optimization function and the need for an a-priori identification of the correct number of factors. Furthermore, additional concerns arise when dealing with compositions, as their bounded structure makes standard statistical tools unavailable. Expressing compositions in centered log-ratio coordinates removes the sum-constraint and preserves the isometry between Euclidean and simplex space, but results in collinear data.

Proceeding from these specifications, two alternative iterative algorithms, ATLD and SWATLD, are considered and compared to Parafac-ALS in a compositional case study. ATLD and SWATLD provide a less affected solution in presence of collinearity and also manage to balance other deficiencies of Parafac-ALS such as convergence speed, presence of swamps and sensitivity to over factoring. Compositional results are properly discussed by focusing on how to explain variation in the array in terms of relative information between compositional parts along a third mode.

FCM 1d: Establishing Dimensionality Cut-off Values For Mixed-Item Format Tests

Doyoung Kim, National Council of State Boards of Nursing (NCSBN); Hong Qian, National Council of State Boards of Nursing (NCSBN); Xiao Luo, National Council of State Boards of Nursing (NCSBN); Ada Woo, National Council of State Boards of Nursing (NCSBN)

The Rasch family models (e.g., one-parameter logistic model and partial credit model) have been used extensively in education, licensure, and certification tests. The Rasch family models assume unidimensionality. The validity of the interpretations of both item and person parameter estimates depends on the degree to which this assumption is supported by a given data. Principal component analysis (PCA) on standardized residuals has been utilized to investigate dimensionality under the Rasch family models. There are several studies recommending cut-off values for PCA on residual correlation matrix (see Smith & Miao, 1994; Smith, 2002; Chou & Wang 2010). However, these studies focused on either the dichotomous or polytomous Rasch model. Considering the growing popularity of mixed-item format tests, it will be valuable to see how the recommending cut-off values work with mixed-item format tests under various conditions. This study proposes to conduct a series of simulations to explore this. The manipulated conditions of the simulations are followed: (I) correlation of two latent factors (0.0, 0.3, 0.5, 0.7, and 1 [null condition]) (II) test length (20, 40, and 60 items), (III) sample size (200, 400, and 800), and (IV) the percentage of polytomous items (10%, 30%, 50%, 70%, and 90%). This study will contribute to the growing literature on checking one of Rasch model assumptions, unidimensionality.

FCM 1e: Investigating Robust Maximum Likelihood Estimators Under Misspecification and Non-Normal Data

Classical maximum likelihood (ML) estimation of structural equation models (SEMs) for continuous outcomes involve normality assumptions, with standard errors (SEs) obtained using the expected information matrix and goodness of fit testing using the likelihood ratio (LR) chi-square statistic. Currently, robust estimation procedures to accommodate the impact non-normal data have become prevalent, where SEs and the chi-squared fit statistic are adjusted. However, choice of information matrix (observed or expected) impacts the calculation of SEs and chi-squared. It is not known which type of information matrix produces optimal results under non-normality and model misspecification.

This study examined three ML-based robust estimation techniques available in Mplus: MLMV, which uses expected information matrix and a mean and variance adjusted LR, MLR-observed with a Yuan-Bentler test statistic and observed information matrix, and MLR-expected (expected information matrix). Data were generated using a two-factor confirmatory factor analysis with varying degrees of correlation among factors (.4, .8, 1.0), and a one factor model was fit to the data. Three levels of non-normality were examined (normal, moderate skew/kurtosis, extreme skew/kurtosis) with sample sizes of 200, 500, or 1000.

Under all sample sizes and distribution conditions, MLR-observed produced trivial (<|5%|) levels of relative bias in SEs. MLM/MV and MLR-expected yielded high levels (> -20%) of bias in SEs when the model was highly misspecified; bias reduced to trivial levels when models were moderately misspecified or correctly specified. Also, MLR-observed yielded more accurate empirical Type-I errors than the alternative procedures. MLR-observed is the recommended procedure to counter misspecification and non-normality.

Thursday, July 14, 2016

Keynote: 8:00 AM - 9:00 AM

Keynote: Bayesian, Fiducial, and Frequentist (BFF): Best Friends Forever? (Chair: Wim van der Linden)
Xiao-Li Meng, Harvard University

Chair: Wim van der Linden

Among paradigms of statistical inferences, Bayesian and Frequentist are most popular, with Fiducial approach being the most controversial. However, there is essentially only one scientifically acceptable way of evaluating any inference method: show me how it performs across replications. And hence the great debate in statistics: Which replications best help us predict real world uncertainty? This unified mode of evaluation provides a prism to reveal the whole spectrum of probabilistic inference foundations. In the familiar Data-Model space, the standard Frequentist's replications fix the Model at the unknown "true" model and let the Data replicate, whereas the Bayesian goes to the other extreme by fixing the Data at the observed and letting the Model vary. The Frequentist thus pays the price of relevance: a method which works on average may not be relevant for the data at hand. In contrast, the Bayesian pays the price of robustness: results are sensitive to prior assumptions about how the Model varies. Fiducial represents one of many possible compromises one can obtain by sliding a ruler along the relevance-robustness spectrum, but it suffers from an incoherent treatment of the Data. Realizing that the differences in inference amount to different choices of replications and there is no one size fits all; Bayes, Fiducial and Frequentism can all thrive under one roof as BFFs (Best Friends Forever) --- only united can we combat the Big Data tsunami.

This talk is based on Liu, K. and Meng, X.-L. (2016). "There is individualized treatment. Why not individualized inference?". Annual Review of Statistics and Its Application, to appear. Available from meng@stat.harvard.edu.

Early Career Award Winner: 9:00 AM - 9:45 AM

Early Career Award Winner: David Magis - Open Source Programming: A New Hope for Psychometric Research
David Magis, University of Liège, Belgium

Chair: Andries van der Ark

Current psychometric research is most often supported by computer software. New research perspectives often imply intensive simulation studies to validate the tested theories or hypotheses, and therefore require accurate, fast and stable implementation. To this regards, open source programming (such as in the R language) is a promising approach allowing for flexible implementation, data generation, replication of studies, and worldwide dissemination. The purpose of this talk is to illustrate how psychometrics and open source programming (with special emphasis on the R language) can interact and contribute to each other, by means of some selected examples. Several topics will be illustrated, among others: why open source programming is (to my opinion) as important as psychometric research; why we need for stable and complete implementation of psychometric and statistical routines for research purposes (for e.g., CAT); how accurate implementation of IRT routines can lead to unexpected theoretical results; why (and how) open source software can be valued as research output. Most examples will arise from the CAT framework and the R package catR for simulating CAT patterns.

Item Response Theory-IRT 5: 10:00 AM - 11:30 AM

IRT 5a: Performances of LOO and WAIC as IRT Model Selection Methods

Yong Luo, National Center for Assessment, Riyadh, Saudi Arabia; Khaleel Al-Harbi, National Center for Assessment

IRT model selection methods such as the likelihood ratio test, AIC, DIC, and cross-validation log-likelihood (CVLL) have been investigated in psychometric literature for both the dichotomous and polytomous IRT models. Another two model selection methods used in Bayesian models, namely leave-one-out cross-validation (LOO) and the widely applicable information criterion (WAIC), despite their theoretical advantages over AIC and DIC, have never been investigated in regard to their performance in selection of IRT models, and it remains unknown how such advantages translate into empirical performances in the context of IRT model selection. In this study, the authors will compare model selection results using the likelihood ratio test, CVLL, AIC, DIC, LOO and WAIC. A simulation study with sample size, test length, and the generating model as manipulated factors will be conducted to examine the type I error and power of those aforementioned model selection methods when selecting from a group of dichotomous IRT models, meanwhile focusing on how LOO and WAIC perform against AIC and DIC. A real data set will also be analyzed to demonstrate the potential inconsistency of using different methods in IRT model selection.

IRT 5b: Examining Effects of Two Adjustments to the lz Person-Fit Statistic

Barth Riley, University of Illinois at Chicago; David Magis, University of Liège, Belgium

The purpose of this talk is to present a simulation study that examined the joint and independent effects of two adjustments to the standardized log-likelihood statistic (lz): (1) correction of the negatively skewed distribution of lz (Snijders, 2001), and (2) improving the sensitivity of the statistic by employing more accurate estimates of item response probability using symmetric functions (Dimitrov & Smith, 2006). Data containing misfitting response patterns were simulated using three aberrant response patterns (cheating, guessing, and inattentiveness), and three levels of aberrance. Non-misfitting responses were generated using the dichotomous Rasch measurement model. Four fit statistics were compared: lz, lz* (Snijders adjustment), lzsym (Dimitrov & Smith adjustment), and lzsym* (both adjustments). Mean Type I error rates were ≤ 0.1 across all conditions. The lz* statistic produced the best control of Type I error, which was often below the nominal Type I error rate, whereas the empirical Type I error rate for the unadjusted lz statistic most closely approximated the nominal rate. In contrast, lzsym and lzsym* yielded empirical Type I error rates larger than the nominal rate, with the discrepancy being particularly pronounced as the length of the test decreased. As might be expected, power to detect misfitting response patterns increased with test length and with the percentage of misfitting response patterns in the sample. Both lzsym and lzsym* evidenced improved power in detecting misfitting response patterns compared to lz and lz*, particularly for guessing response patterns and/or on shorter (i.e., 10 item) tests.

IRT 5c: Performance of the S-X2 Statistic for Higher-Order IRT Models

Xue Zhang, NORTHEAST NORMAL UNIVERSITY; Chun Wang, UNIVERSITY OF MINNESOTA; Jian Tao, NORTHEAST NORMAL UNIVERSITY

Tests of item model misfit are often performed to validate the use of a particular model in item response theory (IRT). As higher-order IRT (HO-IRT) model has been more widely applied, there is a lake of promising solutions to detect the item model misfit in HO-IRT model. Hence investigating item fit for HO-IRT model becomes important. In this study, we extend Orlando and Thissen’s S-X2 item fit index to HO-IRT models. The performance of S-X2 is evaluated in terms of empirical Type I error rates and power, and its performance is compared to McKinley and Mill’s G2. The manipulated factors include test length, sample size, and level of dimension correlation. The results from simulation study demonstrate that the S-X2 is promising for higher-order items.

Symposium 4: Noncognitive Assessment in Education and Workforce Selection: 10:00 AM - 11:30 AM

Symposium 4a: A Multidimensional Forced-Choice IRT Approach to High-Stakes Personality Testing

Stephen Stark, University of South Florida; Oleksandr S. Chernyshenko, Nanyang Technological University; Fritz Drasgow, University of Illinois at Urbana-Champaign

A growing body of research indicates that personality test scores can predict a variety of educational and occupational outcomes (e.g., Salgado & Tauriz, 2014). To deal with response biases, particularly socially desirable responding or "faking good," which is common in high-stakes settings, researchers have expressed interest in multidimensional forced-choice formats as an alternative to ordinal-response (Likert-type) items. Our presentation will describe an item response theory (IRT) approach to constructing and scoring multidimensional forced-choice tests. We will discuss the multi-unidimensional pairwise preference model (MUPP; Stark, 2002; Stark, Chernyshenko, & Drasgow, 2005), methods for MUPP statement and person parameter estimation, a computerized adaptive testing (CAT) algorithm for efficiently selecting pairwise preference items, and indices that were developed to identify various forms of aberrant responding prior to decision making. We will summarize results over the last decade from simulation studies examining the efficacy of these IRT methods, as well as criterion-related validity evidence for a personality assessment, known as TAPAS (Tailored Adaptive Personality Assessment System; Drasgow et al., 2012; Stark et al., 2014), which has been used to screen U.S. military applicants since 2009.

Symposium 4b: Validity of Standardized Non-cognitive Ratings for Graduate School Admissions

David Klieger, Educational Testing Service

Experts in graduate and professional education have encouraged the use of valid assessments of non-cognitive skills in making graduate and professional school admissions decisions. Based on theory and empirical research in other contexts (e.g., the workforce, undergraduate education), they hypothesize that non-cognitive assessments reliably can measure determinants of success in graduate and professional school that would result in smaller or reversed group mean score differences generally observed to be in favor of majority groups on cognitive assessments. However, research has revealed an enduring and prevalent tradeoff between maximizing racial-ethnic diversity and predictive validity in assessment and selection. Therefore, this paper investigates whether non-cognitive assessment can be an exception to this "diversity-validity tradeoff." As a proof of concept, we report findings for the ETS Personal Potential Index (PPI), a standardized assessment by third parties of the non-cognitive skills of applicants to graduate and professional school. Our PPI studies show that a non-cognitive assessment can contribute simultaneously to predictive validity and racial-ethnic diversity in graduate school admissions. Specifically, the PPI predicts cumulative graduate grade point average (CGGPA), provides incremental information over and above U.S. domestic undergraduate grade point average (UGPA) and the GRE General Test (GRE) for predicting CGGPA, and predicts who obtains a CGGPA of at least 3.8. At the same time, in comparison to more cognitive assessments (UGPA and the GRE), use of the PPI resulted in smaller, or reversals in the group mean differences usually favoring the majority group on more cognitive assessments.

Symposium 4c: Predicting Employee Job Performance using Forced-Choice Personality Scores

Jacob Seybert, Educational Testing Service; J.R. Lockwood, Educational Testing Service

As the use of forced-choice personality measures for employment testing continues to grow, there is a need to examine the predictive strength of those scores not only across organizations, but also across languages and cultures. This study investigated the ability of personality dimensions measured using a pairwise preference response format to predict employee job performance ratings. The analysis included nine datasets representing multiple organizations across four language/region combinations. Multiple distinct modeling approaches were used to assess out-of-sample predictive strength of the dimension scores for performance ratings. The approaches included (1) R^2 designed to estimate out-of-sample predictive validity of a linear regression of ratings on 12 dimension scores; (2) cross-validation R^2 computed from linear regression models chosen by stepwise variable selection applied to 100,000 random 80% training samples; and (3) R^2 computed from leave-one-out cross-validation of boosted regression tree models selected by 10-fold cross validation within training samples. These distinct approaches provided highly similar estimates of predictive strength, indicating that the 12 personality dimension scores examined can predict up to 4% of the variance in job performance ratings with a value of about 3% being typical if datasets similar to those in the current study are used for prediction. Additionally, there were interesting trends across the datasets, including that the Achievement dimension consistently showed a positive relationship with performance, and that Sociability often showed a negative relationship. A detailed summary of these results will be presented.

Symposium 4d: Predicting Graduate School Outcomes with a Forced-Choice Personality Test

Patrick Kyllonen, Educational Testing Service

In this study we administered an adaptive forced-choice personality test to 607 graduate school students; we had extensive outcome data on 321 of those (the sample was a mix of ages, race/ethnicities, male-female, 40% foreign national). The test presented statement pairs (e.g., I like to keep a schedule – I love going to meetings) and respondents were instructed to select the statement "more like you." The test comprised 120 pairs measuring 15 dimensions based on the five-factor model of personality (e.g., achievement, sociability, collaboration, curiosity), took approximately 25 minutes to administer, and was scored based on Stark, Drasgow, and Chernyshenko's (2005) multi-unidimensional pairwise preference model following generalized graded unfolding model item calibration.

Simple correlations showed moderate relationships (r > .20) between personality factors and outcomes (e.g., cumulative grade-point average [CGPA], participation in student government, contributions to the community). The strength of relationship between a personality factor composite and CGPA was about the same as between standardized cognitive test scores and CGPA; after the latter correlation was corrected for range restriction due to the study being done on students selected by the cognitive test score rather than applicants. Multiple regression analyses showed that the personality composite added to the prediction of outcomes, notably CGPA, controlling for undergraduate grades and for standardized test scores, and that this relationship held under 5-fold cross-validation. We also found smaller differences between subgroups on the personality measure compared to those found on standardized cognitive test scores.

Estimation and Computational Methods-ECM 2: 10:00 AM - 11:30 AM

ECM 2a: Computationally Efficient Power and Sample Size Determination for Mediation Models

Alexander M. Schoemann, East Carolina University; Aaron J. Boulton, University of North Carolina; Stephen D. Short, College of Charleston, USA

Mediation analyses abound in behavioral research. Current recommendations for assessing power and sample size in mediation models are to use a Monte Carlo power analysis and to test the indirect effect with a bootstrapped confidence interval (e.g. Zhang, 2014). Unfortunately, these methods have rarely been adopted by researchers due to limited software options and the computational time needed. We propose a new method and convenient tools for determining sample size and power in mediation models. The method uses a randomly varying N approach to sample size determination (Schoemann, Miller, Pornprasermanit, & Wu, 2014), where N varies across plausible values for each replication (e.g., N = 200 - 500). For each replication, the sample size and significance of each parameter (0 = not significant, 1 = significant) are recorded. When the simulation is complete, parameter significance is regressed on sample size using logistic regression. For a given sample size, the predicted probability from the logistic regression equation is the power to detect an effect at that sample size. In addition, our method uses Monte Carlo confidence intervals to test the indirect effect, a computationally efficient method that provides accurate tests of indirect effects and complex functions of indirect effects (Tofighi & MacKinnon, 2016). We will demonstrate the accuracy of our new method through Monte Carlo simulations and an easy to use Shiny application that implements our method. These developments will allow researchers to quickly and easily determine power and sample size for simple and complex mediation models.

ECM 2b: Agree to Disagree: Tests for Between-Groups Differences in Within-Group Agreement

Alice M. Brawley, Clemson University; Patrick J. Rosopa, Clemson University; Jamie M. Fynes, Clemson University; Cecelia A. Rosopa, University of South Carolina

Researchers in the behavioral and social sciences typically examine differences between independent means, such as average perceived supervisor support in various teams or average mental workload across type of work schedule. In addition to mean differences, researchers often emphasize the meaning of variability within groups. For example, agreement about organizational climates or variability in stressors can predict important outcomes. To better understand these substantive issues using statistical procedures, we proposed six statistical tests for between-group differences in within-group variability, as indexed by interrater agreement. In a Monte Carlo simulation, we evaluated the Type I error and statistical power of these tests and two existing procedures (Pasisz & Hurtz, 2009) under a number of realistic experimental conditions, such as small group sizes, disproportionate group sizes, and small differences in agreement. Results showed that the Pasisz and Hurtz (2009) F-test provides the best control over Type I error, but this test did not consistently provide the greatest power. The bootstrap resampling technique, approximate randomization test, and Levene's test were more liberal tests, resulting in inflated Type I error and generally increased statistical power. The O'Brien and Welch df adjustments to the F-ratio, the Brown-Forsythe test, and equality of correlations procedures were very conservative tests, often resulting in Type I error rates much lower than the nominal level and relatively low statistical power. Practical recommendations and future research applications in the behavioral, health, and social sciences are discussed.

ECM 2c: Simulating Non-Normal Distributions Using the Normal, Logistic, and Uniform Distributions

Mohan Dev Pant, University of Texas at Arlington; Todd Christopher Headrick, Southern Illinois University-Carbondale

Fleishman's (1978) third-order power method polynomials are generally used for simulating non-normal distributions with specified values of skew (L-skew), kurtosis (L-kurtosis), and Pearson correlation (L-correlation). L-moment (Hosking, 1990) based power method polynomials are superior to their conventional product-moment based counterparts in terms of estimation of parameters and distribution fitting (Headrick, 2011). Although conventional and L-moment based polynomials are capable of producing distributions with large departures from normality, they are extremely peaked and generally not representative of real-world data (Headrick, 2011). To obviate this problem, three families of power method distributions are introduced by mixing third-order power method polynomials based on standard normal, logistic, and uniform distributions using a doubling technique (Morgenthaler & Tukey, 2000). Systems of equations associated with skew (L-skew) and kurtosis (L-kurtosis) are derived for each of the three families. Also developed is a methodology for simulating correlated power method distributions from these three families based on Pearson correlation and L-correlation procedures. The methodology can be applied in a variety of settings such as modeling events and Monte Carlo simulation studies. Further, it is demonstrated that estimates of L-skew, L-kurtosis, and L-correlation are substantially superior to estimates of conventional product-moment based skew, kurtosis, and Pearson correlation in terms of both relative bias and efficiency when distributions with larger departure from normality are involved.

ECM 2d: Constructing UMP Test Based on Exact Distribution for IRT Models

Xiang Liu, Teachers College, Columbia University; Zhuangzhuang Han, Teachers College, Columbia University; Matthew Johnson, Teachers College, Columbia University

In educational and psychological measurement, short test forms are often used. The asymptotic normality of MLE of person parameter in IRT models does not hold under this scenario. As a result, constructing hypothesis testing or confidence interval of the person parameter requires knowing the exact distribution or its approximation. However, the computation involved is often prohibitively expensive. In this paper, we propose a general framework for constructing uniformly the most power test for IRT models within the

exponential family. In addition, an efficient branch and bound algorithm for calculating the exact p-value is introduced. The Type I error rate and statistical power of the proposed exact test is examined through a simulation. We also demonstrate its practical use through analyzing a real data set.

Applications-APP 2: 10:00 AM - 11:30 AM

APP 2a: Detecting Test Results Involved in Test Collusion Incidents

Jiyoon Park, Federation of State Boards of Physical Therapy; Yu Zhang, Federation of State Boards of Physical Therapy; Lorin Mueller, Federation of State Boards of Physical Therapy

Test fraud has recently received increased attention in the field of educational testing. As types and forms of test fraud become diverse, the use of comprehensive analysis is recommended for investigating those incidents. Specifically, test collusion is one of the types that commonly occur in test administrations, especially in high stakes tests. Test collusion arises when test takers share test information with each other, when test information is provided to test takers, or when candidates study together in pairs or groups. Although testing collusion is one of the primary concerns in test administrations, there have not been studies that provide statistical evidence to detect test results as clusters that are potentially involved in test collusion.

In this study, we propose a method for detecting examinee clusters who potentially colluded during test administrations. The detection process entails two stages as an algorithm. In the first stage, pairs with significantly similar responses are identified using similarity indices. In the second stage, individual examinees identified from the first stage are grouped based on examinee characteristics and counts being flagged. Simulated data are used to demonstrate the algorithm.

APP 2b: Comparing Model Estimates for Predicting High School Dropouts.

Justin Long, The University of North Carolina at Greensboro; S. Austin Cavanaugh, The University of North Carolina at Greensboro

A school district in North Carolina explored the effects of a newly developed reading intervention by comparing students receiving intervention with similar students who did not. To develop a comparable control group, propensity scores (PS) were generated for all students attending schools where the intervention took place. Due to student nesting within school and grade concerns were raised about the validity and accuracy of the results of the propensity score matching procedure with the decrease in sample sizes. An alternative method for creating matched samples, called "Brute Force Matching", was employed which generates multiple matched samples based numeric covariates (such as test scores) then identifies samples that are most similar to the categorical variables of the treatment group.

This study will explore the differences between PSM and "Brute Force Matching" methods in two ways. First, in terms of the comparability of the descriptive statistics of the matched groups created by each. Second, by comparing the results of the final analysis of the reading intervention when using one matched group versus the results obtained using the other matched group. Specifically, we will be comparing the intervention group and control group in terms of their mean scores on a reading assessment and the proportional representation of different ethnic groups and genders between intervention and control. Additionally, the two methods will be compared via a simulation study where sample sizes are manipulated (i.e. N= 25, N=50, N=100).

APP 2c: MOOCs from an Item-Centric Perspective

Ben Domingue, Stanford University; Alex Kindel, Stanford University; Frank Rijmen, Association of American Medical Colleges; Andreas Paepcke, Stanford University; Mitchell Stevens, Stanford University

Millions have enrolled in massive open online courses (MOOCs) since 2012. Research thus far has focused on learner behavior in MOOCs while the behavior of items in MOOCs remains poorly understood despite numerous unique features. Prior research suggests that items in MOOCs may be affected by higher levels of item position effects and non-ignorable missingness than typically encountered in educational measurement while also having a number of unique design features. From the perspective of systematically characterizing and improving the content of MOOCs, understanding these issues and how they are related to learner engagement is crucial. Our analysis leverages a dataset describing item views and responses across 25 MOOCs offered by Stanford University between 2013 and 2015. Each MOOC has between 20 and 200 items (mean=95) and between 3700 and 100,000 learners enrolled (mean=22,000). The courses themselves span a wide range of instructional formats, content areas, and enrollment profiles. To begin to understand this variation, we first describe variation in item characteristics (e.g. mean item response correctness) across MOOCs. Subsequently, we describe variation in non-standard factors associated with these items (e.g. response time, number of attempts, placement, semantic load) in an attempt to determine how these factors affect item response and continued student engagement. Finally, we consider psychometric models (e.g., does a consideration of responsive time offer new information about learner ability or engagement compared to item response alone) of MOOC item responses in light of above findings.

APP 2d: Spectral Clustering and its Application to Large Scale Personality Assessments

Yunxiao Chen, Columbia University

Popular self-report personality assessments usually contain large numbers of items and respondents. An important topic in analyzing data from these assessments is to form scales that measure different characters of personality, typically through cluster analysis of items. However, the large scale of data often poses computational challenges for the traditional clustering methods, such as the K-means and the latent class model based algorithms. In this talk, we propose a spectral clustering algorithm for the item cluster analysis. The method is computationally efficient and often outperforms traditional clustering algorithms. In addition, a Monte Carlo based method is proposed for choosing the number of clusters. The proposed method is evaluated through a real dataset from the revised Eysenck Personality Questionnaire.

Measurement Invariance and Differential Item Functioning- DIF 3:
10:00 AM - 11:30 AM

DIF 3a: An Application of Differential Item Functioning Analysis with Rasch Trees

Levent Ertuna, Hacettepe University

This study analyzes if the subtests (Turkish, mathematics, science and social sciences) of Student Achievement Determination Exam (ÖBBS), which was conducted in 2005, shows Differential Item Functioning (DIF) with the use of Rasch Trees Method (according to the combination of gender, educational and socio-economic opportunities of parents variables). It is significant in terms of using more than one variable at once in defining DIF on a high stake test. It is descriptive research which defines an existing situation. While the population of the research consists of 1,285,862 5th graders around Turkey, the sample is composed of 29,889 5th graders who took ÖBBS in 2005. The DIF analysis was made with Rash Trees method and appropriate tree graphics were prepared. However, item parameters were calculated for each node. This method is based on the Item Response Theory, which was developed by Strobl, Kopf and Zeileis (2010), and it is a recursive partitioning technique underlying its structure. This is a global DIF method which displays-for the whole test-how item parameters work for subgroups in the form of tree branches. The results indicate that all the subtests of ÖBBS present DIF according to the aforementioned variables. The Rasch Tree involves branches in various levels, 19 nodes for Turkish and Science tests and 21 nodes for Social Sciences test. Nonetheless, socio-economic opportunities for Turkish and Social Sciences tests and father education for mathematics and science tests is the primary source for DIF. Moreover, for some items, item parameters differentiate in terms of terminal nodes.

DIF 3b: Guideline for Empirical Application of DIF Analysis on Pretest Items

JP Kim, ACT, Inc.; Tianli Li, ACT, Inc.

For many testing programs, differential item functioning (DIF) analysis is conducted using one selected method and flagged items are sent to subject matter experts for fairness review. It is not uncommon that very few items among the flagged items are judged as biased by the experts. Even the most popular DIF analysis method has its limitations under particular conditions, such as the issue of inflated false positives. The purpose of this study is to investigate the advantages of using multiple methods in detecting DIF on pretest items.

The combined use of two commonly adopted DIF analysis methods for multiple-choice items, Mantel-Haenszel (MH) and SIBTEST, was investigated. MH is statistically powerful but it shows inflated Type I error rates when subgroup abilities differ in the matching variable, discrimination of the studied item is high and the matching variable is less reliable. SIBTEST provides better control of Type I error rates under these conditions by adjusting for the unreliability of observed scores, but has relatively low statistical power. Logistic regression and IRT based DIF analysis methods are not considered due to their large sample size requirements which is rare in a pretesting environment.

The manipulation of empirical data will be used to create various conditions in relevant characteristics of testing (subgroup abilities, reliabilities of matching variables). Through the examination of all possible combined outcomes of the two methods, guidelines are expected to be offered for empirical application of multiple methods in DIF analysis on pretest items toward creating a more effective DIF procedure.

DIF 3c: Using Keystroke Log Features to Identify Group Differences

Mo Zhang, Educational Testing Service; Randy Bennett, Educational Testing Service; Paul Deane, Educational Testing Service; Peter van Rijn, Educational Testing Service

Observable information collected during the process of writing composition (as opposed to the text in its final state) has shown to provide evidence for and/or inferences on one's cognitive activities during writing, for which evidence the final text is unable to supply. Such information can be accurately and efficiently collected via keystroke logging, and can include the actions taken in producing the text (e.g., insertion, deletion, pause), the duration of such actions (e.g., inter-key pause burst length), as well as when and where an action occurs in the composition process. In this study, using keystroke log data collected from a middle school summative writing assessment, we investigate whether and how students on different writing proficiency levels differ in their essay writing processes. We apply the methodological principles of differential feature functioning and examine the processes features conditional on certain writing proficiency scores. We also take special interests in examining the differences in writing processes for various demographic groups (e.g., white vs. black; low SES vs. not-low SES). The results of this study can potentially help inform and improve the teaching and learning practices for writing in classrooms.

DIF 3d: Comparison of Mantel-Haenszel and IRT Approaches for Investigating Item Exposure

Weiwei Cui, College Board; Amy Hendrickson, College Board

Test items exposed before an administration date, for example as part of embedded pretesting, may cause test items to perform differently between test takers with and without exposure to these items. This differential exposure, thus, may lead to differential item functioning (DIF). As Thissen (1987) pointed out, DIF is a serious threat to the validity of test scores. Both contingency table based Mantel-Haenszel (MH) and IRT based item drifting analyses can be used to evaluate the effects of potential item exposure. MH approach evaluates the potential item exposure by examining the difference on a test item between test takers with and without exposure to test items. The IRT based item drifting approach evaluates the potential item exposure by examining the item invariance between the test takers with and without exposure to test items. MH approach evaluates DIF on single item level while IRT approach can be used to evaluate DIF on single item or a set of items. This study compares the sensitivity of the two approaches to the degree of item exposure (e.g., the proportion of the exposed items and the proportion of the test takers who exposed to test items) using simulated data.

Network Analysis-NET 1: 10:00 AM - 11:30 AM

NET 1a: Asymmetric Multidimensional Scaling of Cognitive Similarities Among Occupational Categories

Akinori Okada, Tama University; Takuya Hayashi, Nara Women's University

Cognitive similarities among 10 occupational categories were derived from more than 2,017 respondents. The similarity from category j to the other categories is obtained from respondents whose occupations belong to occupational category j. The mean of all obtained similarities from occupational categories j to k is derived to develop a matrix of similarities among occupational categories. The (j,k) element of the matrix is the mean of similarities from occupational categories j to k. The matrix is asymmetric because the similarity from occupational categories j to k is not necessarily equal to that from occupational categories k to j. The asymmetric similarity matrix was analyzed by the asymmetric multidimensional scaling (Okada & Tsurumi, 2012) which shows asymmetric relationships among occupational categories by representing each occupational category as a point in a plane (configuration) for each dimension, where the horizontal axis represents the closeness from each occupational category to the other occupational categories, and the vertical represents the closeness from the other occupational categories to the occupational category. The three-dimensional result was chosen as the solution. The configuration corresponding to Dimension 1 represents the difference between the agricultural occupational category and the other occupational categories, where the relationships are almost symmetric. The configuration corresponding to Dimension 2 represents asymmetric relationships between the female dominated to the male dominated occupational categories. Lastly, the configuration corresponding to Dimension 3 represents asymmetric relationships between the more autonomous or prestigious and less autonomous or prestigious occupational categories.

NET 1b: Exploratory Graph Analysis: A New Approach for Estimating Dimensionality

Hudson F. Golino, Universidade Salgado de Oliveira; Sacha Epskamp, University of Amsterdam

The estimation of the correct number of dimensions is a long-standing problem in psychometrics. Several methods have been proposed, such as parallel analysis (PA), multiple average partial procedure (MAP), the maximum-likelihood approaches that use fit indexes as BIC and EBIC, and the less used and studied approach called very simple structure (VSS). In the present presentation a new approach to estimate the number of dimensions will be introduced and compared via simulation to the traditional techniques pointed above. The approach proposed in the current paper is called exploratory graph analysis (EGA), since it is based on the graphical lasso with the regularization parameter specified using EBIC. The number of dimensions is verified using the walktrap, a random walk algorithm used to identify communities in networks. In total, 3,200 data sets were simulated to fit known factor structures, with the data sets varying across different criteria: number of factors (2 and 4), with five dichotomous items each, sample size (100, 500, 1000 and 5000) and correlation between factors (orthogonal, .20, .50 and .70), resulting in 32 different conditions. For each condition, 100 data sets were simulated using lavaan. The result shows that the EGA was the only technique able to correctly estimate the number of dimensions in the four-factor structure when the correlation between factors were .7, showing an accuracy of 61% for a sample size of 1,000, and 100% for a sample size of 5,000 observations, against an accuracy of 0 for all the other methods, in all conditions studied.

NET 1c: Identifying the Presence and Number of Clusters/Communities in Networks

Michaela Hoffman, University of Missouri-Columbia; Douglas Steinley, University of Missouri-Columbia

Detecting groups in network analysis (e.g., community detection) is a type of cluster analysis where the goal is to identify homogeneous subgroups within a given data set. Typically, the true number of clusters is unknown and estimating the correct number is one of the most difficult problems in cluster analysis -- this difficulty holds when trying to detect so-called communities in network data. We present a method for both

determining the number of communities and assessing the strength of the communities. By calculating a similarity measure between each pair of observations in the network, we obtain a distribution of agreement/disagreement scores. This distribution is then subjected to a simple test for bimodality that can be used to identify the presence/absence of communities. Finally, the expected value and variance of the distribution is derived to estimate the number of clusters and provide a measure of uncertainty for the results.

NET 1d: Not Sure About Linearity? Boost Multivariate Outcomes with Trees!

Patrick Miller, University of Notre Dame; Gitta Lubke, University of Notre Dame; VU University Amsterdam

When the number of predictors is large, finding relevant predictors with potentially non-linear effects can be challenging. This challenge is made more difficult when considering that in psychological research outcomes are commonly measured multivariately. Model based recursive partitioning procedures are useful but require strong assumptions, can be difficult to interpret, and may not directly answer the research question of interest. To address these problems, we introduce a multivariate extension to boosted decision trees (Friedman, 2001) called multivariate tree boosting. Boosting decision trees (Friedman, 2001) flexibly approximates non-linear effects and interactions among many predictors without the need to a priori specify any effects, and can be easily estimated with minimal tuning of meta-parameters.

We illustrate several ways to interpret a resulting model, select important variables, visualize non-linear effects, detect possible interactions, and identify predictors that cause two or more outcome variables to covary. These procedures can be used to select both important predictors and outcomes. Estimation, tuning, and interpretation of the the model can be easily accomplished using our R package ‘mvtboost’. Simulations verify that our approach identifies predictors and achieves good prediction performance when predictors have non-linear effects compared to (penalized) multivariate multiple regression and multivariate decision forests. In addition we will discuss ongoing work on extensions of multivariate tree boosting to account for non-independence of observations in multilevel and longitudinal settings.

NET 1e: A New Badness-of-Fit Function for Nonmetric Multidimensional Unfolding: Stress-3

Frank Busing, Leiden University; Patrick Groenen, Erasmus University Rotterdam; Willem Heiser, Leiden University

Almost from conception, nonmetric multidimensional unfolding suffered from the degeneracy problem. Irrespective of the data, stress-based unfolding algorithms invariably converge into perfect, but useless solutions. The uselessness of the solution is characterized by equal and constant distances between the points representing judges and stimuli, while at the same time providing a global minimum for the loss function. Early attempts to overcome the problem used clever normalizations of the standard loss function such as the variance of the (pseudo-)distances (Kruskal, 1964, 1965), in order to steer away from such a trivial solution, but none was unambiguously successful. Busing, Groenen, and Heiser (2005) conjectured that the variance might not be the proper statistic to use as normalization factor and that the normalization or penalty factor might simply not be strong enough. Their complicated loss function, called penalized Stress, uses the coefficient of variation as a penalty factor and contains two so-called penalty parameters to (fine) tune the penalty factor. Here, we will introduce a much simpler loss function, without the additional penalty parameters, but capable of avoiding trivial solutions called Stress-3. Stress-3 can be seen as a continuation of the scale factor series introduced by Kruskal and uses the aforementioned coefficient of variation for this purpose. We will describe the new loss function, discuss several features, compare Stress-3 with both Stress-2 and penalized Stress, and use well-known data sets to demonstrate the benefits of the new loss function.

Validity and Reliability-VAL 2: 10:00 AM - 11:30 AM

VAL 2a: Robust Estimation to Instrumental Variables in Two-Stage-Least-Squares Model

Dingjing Shi, University of Virginia; Xin Tong, Department of Psychology, University of Virginia

One challenge in causal inference is that the treatment assignment and treatment delivery can be unmatched (i.e., participants being assigned to the intervention do not actually do the intervention). A common solution is to select instrumental variables to estimate the local average treatment effect. When both the treatment and instrument variables are normally distributed, model parameter estimates are unbiased. However, practical data usually violates normality assumptions. For example, in a study to investigate the effectiveness of financial incentives (treatment) to retain teachers in low-performing schools (outcome), only teachers meeting certain criteria are eligible (instrument) to be considered for the financial assistance (Steel et al., 2010). The potential non-normal data may lead to biased estimates of the treatment effect. The purpose of this study is to propose a robust estimation method to model the non-normal heavy tailed data with instrumental variables in the framework of two-stage-least-squares (2SLS) model. Student's t-distribution is applied to account for the heavy tails of the data at the two stages. Particularly, first-stage residuals are assumed to follow a t-distribution if the instrument-affected treatment data are non-normal; similarly, second-stage residuals follow the t-distribution with non-normal outcome data. For real situations, the non-normality can occur at either or both stage(s), and thus there are three types of robust distributional models. In this study, the performance of the proposed robust models is evaluated using Monte Carlo simulations, and compared to that of the ordinary 2SLS model with normality assumptions at both stages.

VAL 2b: Incremental Validity is a Statistically Problematic Concept

Jacob Westfall, University of Texas at Austin; Tal Yarkoni, University of Texas at Austin

Social scientists often seek to demonstrate that variation in some construct (e.g., verbal ability) can predict variation in some meaningful outcome (e.g., college GPA), over and above other related constructs (e.g., socio-economic status). In other words, they wish to establish that the focal construct has “incremental validity.” However, these claims are typically supported by measurement-level regression models that fail to consider the (un)reliability of the observed scores. It is fairly well known to psychometricians (but not well known to others) that attempts to establish incremental validity using multiple regression will exhibit inflated Type 1 error rates. But even the initiated may not fully appreciate the scope and magnitude of the problem in published research. We first revisit the classic incremental validity argument. We show analytically that Type 1 error rates can be extremely high under parameter regimes common in many psychological domains. Error rates are highest when reliability is moderate (.4 to .5), and can approach 100% surprisingly quickly as the sample size grows. Next, we discuss two common variants of the classic incremental validity argument—which we call the argument for separable constructs and the argument for improved measurement—that are not usually recognized as such, and show that they share the same magnitude of statistical problems. Finally, we show that although SEM-based approaches can be used to account for measurement error and control the Type 1 error rate, these methods can be severely underpowered for answering incremental validity questions, often requiring many hundreds or thousands of participants.

VAL 2c: Potential Test Information for Multidimensional Tests

Katherine Jonas, University of Iowa

Test selection is guided by test information. Most existing formulations of test information are specific to the sample for which they are estimated, with the result that test information will vary from sample to sample. Recently, measures of test information have been developed that quantify the potential informativeness of the test, a quantity that is defined by the properties of the test, and which is independent of the characteristics of the examinees. As of yet, however, measures of potential information have only been applied to unidimensional tests. In practice, psychological tests are often multidimensional. Furthermore, multidimensional tests are often used to estimate one trait among many others, which are then nuisance traits. This talk describes measures of potential test information for multidimensional tests, as well as measures of marginal test information--that is, test information with regard to one trait within a multidimensional test. Performance of the metrics is tested in simulated and observed data. The measures allow for the direct comparison of two multidimensional tests that assess the same trait, facilitating test selection, precision, and validity.

VAL 2d: Evolution of Psychometric Skills for Developing and Evaluating Psychological Tests

Kurt F. Geisinger, Buros Center for Testing, University of Nebraska-Lincoln

This paper will trace the evolution of psychometric skills needed to develop and/or evaluate psychological tests and measures. Specific mention will be made of classical psychometrics, especially reliability; exploratory factor analysis; generalizability theory; confirmatory factor analysis; item-response theory; equating methodology; Bayesian statistics; and changing conceptions of validity. The presentation will describe impact of software improvements on reliability, item analysis, and equating. It will trace trends in the Buros reviews of tests showing that many who evaluate tests have not kept up with the conceptual and methodological changes. Implications for the training both of psychometricians and psychologists more generally will be discussed.

Computer-based Testing- CBT 3: 1:30 PM - 3:00 PM

CBT 3a: Assessing Content Balancing: CAT vs MST

Halil Ibrahim Sari, University of Florida; Anne Corinne Huggins-Manley, University of Florida

Many comparison studies have been conducted to investigate efficiency of the statistical procedures across CAT and MST. Although it is directly related to the validity, score interpretation and test fairness, non-statistical issues of adaptive tests, such as content balancing, haven't been given more attention. It's consistently asserted in several studies a major advantage of MST is that it controls for content better than CAT. Yet, the literature doesn't contain a study that specifically compares CAT with MST under varying levels of content constraints to verify this claim. A simulation study was conducted to explore how accurate outcomes these two adaptive tests produce when content balancing procedures are strictly met. One CAT and two MST designs (1-3 and 1-3-3 panel designs) were compared across several manipulated conditions including total test length (24-item and 48-item test length) and number of controlled content area. The five levels of content area condition include zero (no content control), two, four, six and eight content area. All manipulated conditions within CAT and MST were fully crossed with one another. This resulted in 2x5=10 CAT (test length x content area), and 2x5x2=20 MST conditions (test length x content area x MST design), for 30 total conditions. 4000 examinees were generated from N(0,1). All other conditions such as IRT model, exposure rate were fixed across the CAT and MSTs. Preliminary results show that the two types of MST designs resulted in similar measurement accuracy for theta estimates but better than CAT.

CBT 3b: Evaluating the Efficiency and Accuracy of an Adaptive Curriculum-Based Measurement

Jennifer Grossman, Teachers College, Columbia University; Young-Sun Lee, Teachers College, Columbia University

Curriculum-based measurement (CBM) consists of short, standardized assessments to evaluate student performance across numerous administrations. CBM effectiveness is affected by the accuracy of the data collected, which is of concern when assessing students of the lowest abilities or who are unmotivated to give their best effort. Teachers may be reluctant to use it due to the time burden associated with multiple administrations. Problems associated with CBM may be remedied by presenting items that match students' ability levels, which may be achieved by developing a computer adaptive test (CAT) version.

A CAT version of a linear computer-based CBM was developed to assess its accuracy and efficiency compared to the linear version. Postsim 3 (Assessment Systems Corporation, 2009), which uses examinees' actual responses to a linear test to estimate performance on an adaptive version, was used. Descriptive statistics, test information functions, Pearson product-moment correlations, bias and root mean squared error estimates were calculated to evaluate the efficiency and effectiveness of the CAT. The CAT version of the CBM yielded more precise estimates of ability than the linear version, but efficiency varied depending on examinee ability level. Implications of these findings are discussed.

CBT 3c: The Continuous G-DINA Model and the Jensen-Shannon Divergence

Nathan Minchen, Rutgers, The State University of New Jersey; Jimmy de la Torre, The University of Hong Kong

Interest in diagnostic assessment has grown rapidly in recent years, as the public has increasingly looked to assessments to enhance educational outcomes. As a result, cognitive diagnosis models (CDMs) have been a popular subject for psychometric researchers. The goal of this research is to expand the methodological toolbox for CDMs by offering two new developments. First, we introduce a generalization of a recently proposed CDM, the continuous-DINA (C-DINA; Minchen, de la Torre, & Liu, under review) model. These models handle continuous response data, rather than binary or polytomous data, allowing for CDMs to be estimated from different kinds of data, such as response times or probability testing. Second, we adapt the Jensen-Shannon Divergence (JSD; Lin, 1991), which is based on the Shannon Entropy and is a measure of quantifying the divergence between two or more probability distributions, for use as an item selection algorithm in cognitive diagnostic computerized adaptive testing. Existing indices, such as those proposed in Kaplan, de la Torre, & Barrada (2015) are insufficient for various reasons. Finally, we conduct two simulation studies. The first establishes the viability of the C-G-DINA model by determining the robustness of model parameter estimation to a variety of different conditions. The second is designed to show the extent to which the JSD selection index provides a substantial improvement over random item selection in terms of examinee classification and test length (for variable stopping rule conditions) across a range of conditions.

CBT 3d: Attribute-level Item Selection Method for DCM-CAT

Yu Bao, University of Georgia; Laine Bradshaw, University of Georgia

Diagnostic classification models (DCMs) are a class of multidimensional statistical models that classify students according to mastery levels of categorical latent traits called attributes. Computerized adaptive testing (CAT) is a testing delivery method that when used in the context of DCMs enables more efficient and accurate classification by strategically selecting items based upon attribute posterior estimates that are updated after each item administration.

Most item selection methods for the DCM-CATs are based on statistical information indices that calculate item information for each attribute profile. Using these attribute profile-level indices, highly certain classifications of subsets of attributes in a profile may mask uncertain classifications of other attributes in the profile. This study develops a new item selection rule based upon information indices calculated for each marginal attribute. These indices are expected to allow the DCM-CAT to more flexibly select items to ensure classification for each attribute within a profile is clear before terminating the DCM-CAT algorithm. Using a simulation study, we investigate the effectiveness of the new item selection method and compare the attribute-level item selection method with the existing attribute profile-level item selection method.

Item Response Theory- IRT 6: 1:30 PM - 3:00 PM

IRT 6a: Using Bayesian Approach to Examine Guessing Based on Students' Ability in 2PL IRT Model

Jiaqi Zhang, University of Cincinnati; Lihshing Leigh Wang, University of Cincinnati; Hok Mark Lai, University of Cincinnati; Christopher M. Swaboda, University of Cincinnati

Although the item response theory (IRT) have been around for the past decades, some basic questions related to the basic measurement models remain unanswered. By introducing a guessing parameter in the 2PL model, guessing is explicitly distinguished in 3PL model. The value of ci represents the lowest probability of the item response function (IRF). The 3PL model assumes that guessing is an item parameter and is the same for all student. However, this assumption is not easily evaluated and unrealistic. In this simulation study, we examined guessing as a student characteristic rather than an item parameter in traditional three parameter logistic (3PL) model for multiple-choice items. Given the large amount of parameters to be estimated, the Bayesian approach is used. Rather than assuming guessing as an item parameter and be the same across all examinees in traditional 3PL model, we assume that guessing is actually based on students' latent ability and model guessing as a student characteristic. Results indicated that when guessing parameter is near 0.5, the point estimates of guessing tend to be overestimated if the guessing parameter is less than .50 whereas that tends to be underestimated if such parameter is greater than .50. The overall bias ranged from −9.02% to 7.15% across all simulated situation.

Keywords: Bayesian IRT, 2PL Model, Guessing, Item Response Function

Although the item response theory (IRT) have been around for the past decades, some basic questions related to the basic measurement models remain unanswered. By introducing a guessing parameter in the 2PL model, guessing is explicitly distinguished in 3PL model. The value of ci represents the lowest probability of the item response function (IRF). The 3PL model assumes that guessing is an item parameter and is the same for all student. However, this assumption is not easily evaluated and unrealistic. In this simulation study, we examined guessing as a student characteristic rather than an item parameter in traditional three parameter logistic (3PL) model for multiple-choice items. Given the large amount of parameters to be estimated, the Bayesian approach is used. Rather than assuming guessing as an item parameter and be the same across all examinees in traditional 3PL model, we assume that guessing is actually based on students' latent ability and model guessing as a student characteristic. Results indicated that when guessing parameter is near 0.5, the point estimates of guessing tend to be overestimated if the guessing parameter is less than .50 whereas that tends to be underestimated if such parameter is greater than .50. The overall bias ranged from −9.02% to 7.15% across all simulated situation.

Keywords: Bayesian IRT, 2PL Model, Guessing, Item Response Function

Although the item response theory (IRT) have been around for the past decades, some basic questions related to the basic measurement models remain unanswered. By introducing a guessing parameter in the 2PL model, guessing is explicitly distinguished in 3PL model. The value of ci represents the lowest probability of the item response function (IRF). The 3PL model assumes that guessing is an item parameter and is the same for all student. However, this assumption is not easily evaluated and unrealistic. In this simulation study, we examined guessing as a student characteristic rather than an item parameter in traditional three parameter logistic (3PL) model for multiple-choice items. Given the large amount of parameters to be estimated, the Bayesian approach is used. Rather than assuming guessing as an item parameter and be the same across all examinees in traditional 3PL model, we assume that guessing is actually based on students' latent ability and model guessing as a student characteristic. Results indicated that when guessing parameter is near 0.5, the point estimates of guessing tend to be overestimated if the guessing parameter is less than .50 whereas that tends to be underestimated if such parameter is greater than .50. The overall bias ranged from −9.02% to 7.15% across all simulated situation.

Keywords:Although the item response theory (IRT) have been around for the past decades, some basic questions related to the basic measurement models remain unanswered. By introducing a guessing parameter in the 2PL model, guessing is explicitly distinguished in 3PL model. The value of ci represents the lowest probability of the item response function (IRF). The 3PL model assumes that guessing is an item parameter and is the same for all student. However, this assumption is not easily evaluated and unrealistic. In this simulation study, we examined guessing as a student characteristic rather than an item parameter in traditional three parameter logistic (3PL) model for multiple-choice items. Given the large amount of parameters to be estimated, the Bayesian approach is used. Rather than assuming guessing as an item parameter and be the same across all examinees in traditional 3PL model, we assume that guessing is actually based on students' latent ability and model guessing as a student characteristic. Results indicated that when guessing parameter is near 0.5, the point estimates of guessing tend to be overestimated if the guessing parameter is less than .50 whereas that tends to be underestimated if such parameter is greater than .50. The overall bias ranged from −9.02% to 7.15% across all simulated situation.

Keywords: Bayesian IRT, 2PL Model, Guessing, Item Response Function Bayesian IRT, 2PL Model, Guessing, Item Response Function

IRT 6b: A Multi-Group Cross-Classified Testlet Model for Mixed-Format Tests

Dandan Liao, University of Maryland-College Park; Hong Jiao, University of Maryland

The key assumption of local item independence and local person independence is required when applying standard item response theory (IRT) models. In practice, some assessments might use items that are clustered around the same content area or scenario, which introduces local item dependence. Other assessments might intentionally or unintentionally involve examinees that are clustered in educational units, which introduces local person dependence. On the other hand, more and more large-scale testing programs and state assessments have adopted mixed-format tests in which multiple choice and constructed response items are administered simultaneously.

Multi-group testlet models have been proposed to account for local item and person dependence concurrently for dichotomously scored items (Jeon et al., 2013). In addition, when the cross-classified structure is fitted with hierarchical structure, the standard error estimates associated with the incorrectly modeled clustering variable will be underestimated (Meyers & Beretvas, 2006). This study intends to generalize multi-group testlet model to accommodate complex cross-classified item clustering structure in testlet-based mixed-format assessments. A multi-group cross-classified testlet model is proposed and model parameter recovery is investigated in simulated study conditions. A real dataset will be analyzed to illustrate the use of the proposed model. This study will provide practitioners with more empirical evidence in analyzing data with cross-classified testlet structures in a multi-group framework. Results of this study could also be helpful for comparing how items function differently across non-equivalent groups when the measurement invariance assumption is violated.

IRT 6c: Validating an IRT-based RS Approach Using Simulated and Empirical Data

Lale Khorramdel, Educational Testing Service; Artur Pokropek, Instytut Badań Edukacyjnych (IBE); Matthias von Davier, Educational Testing Service

Personality constructs, attitudes and other noncognitive variables are often measured using rating or Likert-type scales which does not come without problems. Respondents who are not motivated, experience fatigue effects, or have problems understanding the questions can give invalid responses in form of response styles that harm the validity and comparability of the measurement. A recent IRT approach to detect response styles in rating data is validated with a simulation study and using empirical data. The IRT approach was introduced by Böckenholt (2012) and is based on a decomposition of rating data into binary pseudo items (BPIs), which are examined using item response theory (IRT) models. The current paper follows the procedure described by von Davier and Khorramdel (2013) and Khorramdel and von Davier (2014) and decomposes responses to a 5-point rating scale into multiple BPIs to examine the extreme and midpoint response style. In addition to the approach described by Böckenholt, which assumes all types of response styles at once, models are presented to differentiate between mid-point and extreme response style. In a simulation study, different scenarios, levels, and consistencies of extreme and midpoint response styles are simulated, and the IRT approach is applied to all resulting data sets. In an empirical study, the approach is further applied to selected scales of the Program for International Student Assessment (PISA) student questionnaire. IRT analyses based on the 2PLM show that the approach and its extension are useful and valid tools to detect and correct for response styles in rating data.

IRT 6d: Detection of Response Shifts in Pretest-Posttest Designs

Wilco H. M. Emons, Tilburg University

The validity of change scores obtained in pretest-posttest designs is questionable when respondents interpret the items differently at pretest and posttest. For example, at posttest a respondent may perceive the response option “often being unhappy" to represent levels of unhappiness that are different than levels perceived at pretest. This is an example of so called response shift, which is caused by recalibration of the item response option at posttest. Response shifts also occur when respondents' fundamental understanding and definition of a latent attribute changes between measurement occasions. For example, respondents may perceive symptoms of distress as an indication of anxiety at pretest but the therapy they undergo may have focused on recognizing different types of stressors, thus leading the measurement away from anxiety at posttest. In this presentation, I will discuss novel IRT methods for detecting response shifts at the group level and the individual level. The methods include approaches based on multiple-group IRT analysis and person-fit analysis. The statistical properties of these methods are investigated using simulation studies. Empirical results are given for the Dutch version of the Outcome Questionnaire-45 (OQ-45) using a sample of N=540 Dutch outpatients.

Structural Equation Modeling- SEM 3: 1:30 PM - 3:00 PM

SEM 3a: Robustness Conditions for Structural Misspecifications in Structural Equation Models

Zachary Fisher, The University of North Carolina at Chapel Hill; Kenneth Bollen, The University of North Carolina at Chapel Hill; Katherine Gates, The University of North Carolina at Chapel Hill

Attention to the consequences of structural misspecifications on parameter estimates in Structural Equation Models (SEMs) is out of sync with the frequency with which such misspecifications occur. It would be helpful for researchers to know whether, for example, problems with the latent variable model ("structural model") spillover into the estimation of the measurement model and vice versa. This paper explores the conditions under which the Model Implied Instrumental Variable – Two Stage Least Squares (MIIV-2SLS) estimator is robust to omitted variables and other structural misspecifications. Bollen (2001) gives general robustness conditions for MIIV-2SLS but does not describe their implications for the two major components of SEMs, the latent variable and the measurement models. Our goal is to explore whether and when structural misspecifications in the latent variable model affect the estimates of the parameters from the measurement model. We prove that if the measurement model is correct and estimated with MIIV-2SLS, it is robust to structural misspecifications of the latent variable model. We also explore conditions under which structural misspecifications in the latent variable model affect the MIIV-2SLS estimates of the measurement model. We illustrate the results with an example and discuss their more general implications.

SEM 3b: Random Coefficient Meta‐Analytic Structural Equation Modeling: A Simulation Study

Zijun Ke, Sun Yat-Sen University; Qian Zhang, College of Education, Florida State University; Xin Tong, Department of Psychology, University of Virginia

Meta-analytic structural equation modeling (MASEM) is a combination of multivariate meta-analysis and structural equation modeling. It is useful for synthesizing multivariate relations among variables from various studies. With multiple studies, however, current approaches to MASEM are not able to capture variations across studies, let alone identifying factors causing between-study variations. Another limitation of current approaches to MASEM is that current correction methods do not perform well for correcting artifacts due to sampling errors in the presence of high missing rates. The main objectives of the current study are a) to propose a random coefficient approach to MASEM (RC-MASEM) to statistically model and explain between-study variations in model parameters using Bayesian analysis; b) to present a new correction method that can appropriately correct artifacts due to sampling error in the presence of high missing rates; and c) to evaluate the performance of the proposed method using a simulation study. Factors considered in the simulation include sample size, number of studies, effect size, and missing rate. Simulation results show that Bayesian estimation of RC-MASEM works satisfactorily in model parameter estimation both for fixed and random effects. The presentation ends with a discussion on practical considerations.

SEM 3c: A Comparative Study of SEM Software: LISREL and Lavaan

Robyn Pitts, University of North Carolina Greensboro; Juanita Hicks, The University of North Carolina at Greensboro

Structural Equation Modeling (SEM) is a set of diverse statistical methods used to describe associations between multiple variables, both latent and observed. Some of the popular software packages for SEM include AMOS, EQS, Mplus, and LISREL, with LISREL historically considered the most widely used. Researchers, teachers, statisticians, and other professionals are unable to utilize these commercial software without paying for licensed access (except in cases in which limited student versions are available for use, in which the number of variables permitted in the SEM analysis is typically limited). By contrast, R is a programming language and statistical software that serves as an adaptable, open-source alternative. The lavaan (latent variable analysis) package in R provides free, commercial-quality software for SEM analyses, including the ability to analyze multiple latent variable models through confirmatory factor analysis, structural equation, latent growth curve, longitudinal, multilevel, item response, and missing data models (Rosseel, 2012). This research study seeks to replicate examples from a widely used SEM textbook (Kline, 2011), in LISREL and in lavaan for comparative study. Differences in results generated by the software for each example are identified and analyzed in context.

Estimation and Computational Methods- ECM 3: 1:30 PM - 3:00 PM

ECM 3a: Stochastic Approximation of the Multidimensional Generalized Graded Unfolding Model

David R. King, Georgia Institute of Technology; James S. Roberts, Georgia Institute of Technology

The multidimensional generalized graded unfolding model (MGGUM; Roberts & Shim, 2010) is a distance-based, unfolding multidimensional item response theory (MIRT) model for measuring person and item characteristics from graded or binary disagree-agree responses to Thurstone or Likert style questionnaire items. It can also be used when graded satisfaction or preference responses to more general multidimensional stimuli (e.g., photographs of faces, samples of coffee, etc.) are obtained. The item parameters in the MGGUM have been estimated using Markov chain Monte Carlo (MCMC; Roberts & Shim, 2010) and marginal maximum a posteriori (MMAP; Thompson, 2014) procedures, although neither procedure efficiently estimates higher-dimensional MGGUMs. An efficient estimation procedure for the MGGUM will increase the utility of the model for applied researchers. One candidate procedure is the Metropolis-Hastings Robbins-Monro (MH-RM; Cai, 2010a; Cai, 2010b; Cai, 2010c) algorithm, which has been shown to efficiently estimate other multidimensional models. This study examines the utility of the MH-RM algorithm for estimating the item parameters in the MGGUM. Data demands, estimation accuracy, and runtime efficiency for the MH-RM procedure are assessed through a parameter recovery study. Furthermore, the practical utility of the procedure is explored through a real data analysis of facial affect responses. Initial results suggest that the algorithm is fast and the estimation accuracy is comparable to that obtained from the MMAP procedure. This study contributes to the development of the MGGUM and to the utility of the model for applied measurement problems.

ECM 3b: High-Performance Psychometrics -- The Parallel-E Parallel-M Algorithm for Generalized Latent Variable Models

Matthias von Davier, Educational Testing Service

This paper presents results on a parallel implementation of the expectation-maximization (EM) algorithm for item multidimensional latent variable models. The developments presented here are based on code that parallelizes both the E step and the M step (the PEPM algorithm). Examples presented in this paper include item response theory, diagnostic classification models, multitrait-multimethod (MTMM) models, and discrete mixture distribution models. These types of models are frequently applied to the analysis of multidimensional responses of test takers to a set of items, for example, in the context of proficiency testing. The algorithm presented here is based on a direct implementation of massive parallelism using a paradigm that allows the distribution of work among a number of processor cores. Modern desktop computers as well as many laptops are using processors that contain two to four cores and potentially twice the number of virtual cores. Many servers use two, four or more multicore central processing units (CPUs), which brings the number of cores to 8, 12, 32, or even 64 or more. The algorithm presented here scales the time reduction in the most calculation-intense part of the program almost for some problems, which means that a server with 32 physical cores executes the parallel E step algorithm up to 24 times faster than a single core computer or the equivalent nonparallel algorithm. The overall gain (including parts of the program that cannot be executed in parallel) can reach a reduction in time by a factor of 6 or more for a 12-core machine. The basic approach is to utilize the architecture of modern CPUs, which often involves the design of processors with multiple cores that can run programs simultaneously. The use of this type of architecture for algorithms that produce posterior moments has straightforward appeal: The calculations conducted for each respondent or each distinct response pattern can be split up into simultaneous calculations.

Key words: Parallel programming, EM algorithm, high-performance computation (HPC), efficient estimation, modern psychometric models.

ECM 3c: Fixed Item Parameter Calibration with MMLE-EM Using a Fixed Prior

Sung-Hyuck Lee, ACT, Inc.; Hongwook Suh, ACT

When new items are administered to a group with old items whose item statistics have been established, prior distributions for the Marginal Maximum Likelihood Estimation with EM algorithm (MMLE-EM) take an important role in estimating their item parameters. The importance of prior distributions becomes more salient in the Fixed Item Parameter Calibration (FIPC) since it affects both the parameter estimation and the scale transformation of the new items. Therefore, when prior distributions are incorrectly specified, the results of the item calibration with the FIPC might be profoundly corrupted. In this study, a new FIPC method is proposed. Unlike the FIPC methods that keep updating a prior during the EM cycles (Kim, 2006) the new FIPC uses a fixed prior that is estimated (Mislevy, 1984) based on the parameters of old items and the responses to them only before the EM begins. Consequentially, only the parameters of new items are updated during the iterative EM cycles. The main advantage of the FIPC with a fixed prior is that poor new items cannot contaminate the calibration results since they are excluded from computing the prior for the FIPC. Accordingly, it is expected that the new FIPC method performs better than the existing FIPC methods which use the responses of a group to both old items and new items in updating the prior when misfit or multidimensionality is assumed for new items. In the presentation, the calibration results of the new FIPC method will be compared with the existing FIPC methods.

ECM 3d: Pseudo-likelihood Estimation of Multidimensional Response Models: Polytomous and Dichotomous Items

Youngshil Paek, University of Illinois at Urbana-Champaign; Carolyn J. Anderson, University of Illinois at Urbana-Champaign

Log-multiplicative association (LMA) models, special cases of log-linear models, can be used as multidimensional item response models for polytomous and/or dichotomous items (Anderson, Verkuilen & Peyton, 2010; Anderson, 2013; Hessen, 2012). A bottleneck to more wide-spread use of LMA models is their estimation for moderate to large numbers of items. Maximum likelihood estimation (MLE) of LMA models requires iteratively computing fitted values for all possible response patterns, which increases exponentially as the number of items and response options per item increase. Anderson, Li and Vermunt (2007) proposed a partial solution using pseudo-likelihood estimation; however, it only applies to models where category scores are known (i.e., models in the Rasch family). In this talk, we propose a straight forward iterative two step algorithm that permits estimation of LMA models where category scores are estimated (i.e., slope parameters in the response nominal and multidimensional generalizations). Based on the results of simulation studies, the parameter estimates from the new algorithm are nearly equivalent to MLE of LMA parameters, works for large numbers of items in a short amount of time, is insensitive to starting values, and converges in a small number of iterations.

Symposium 5: Statistical Consulting: 1:30 PM - 3:00 PM

Symposium 5a: Statistical Consulting

Carolin Strobl, University of Zurich

Many of us offer statistical or psychometric consulting to applied researchers or practitioners in- and outside of academia - whether informally or as a part of our job description. Consulting teaches us important lessons from practice and can be very rewarding, but it can also be very time consuming and expectations may differ between us and our clients. On the panel, five researchers with extensive experience in statistical consulting will give short presentations about how they organize their consulting services, what challenges they have encountered and how they are coping with them. After this round of introductions, the panel is open to questions and discussion with the audience.

Symposium 5b: Statistical Consulting

Ric Luecht, University of North Carolina at Greensboro

Abstract is not available at this time

Symposium 5c: Statistical Consulting

Wim van der Linden, Pacific Metrics Corporation

Abstract is not available at this time

Symposium 5d: Statistical Consulting

Andries van der Ark, University of Amsterdam

Abstract is not available at this time

Symposium 5e: Statistical Consulting

Dan Bolt, University of Wisconsin-Madison

Abstract is not available at this time

Measurement Invariance and Differential Item Functioning- DIF 4: 1:30 PM - 3:00 PM

DIF 4a: On the use of ROC Curves in DIF Simulation Studies

David Magis, University of Liège, Belgium; Francis Tuerlinckx, KU Leuven, Belgium

Simulation studies are often used to compare methods to detect differential item functioning (DIF). However, comparing the performance of such methods can become complicated when the identification of DIF items relies on statistics based on pre-defined significance level or on pre-established cutoff values. DIF methods based on conceptually different approaches may therefore become incomparable in terms of summary DIF statistics such as false alarm rate or hit rate. The purpose of this talk is to overcome this analytic issue by introducing receiver operating characteristic (ROC) curves in this context. ROC curves allow for global comparison of methods' performances by computing pairs of (false alarm, hit) rates and representing them on a common scatter plot. Several summary ROC statistics can be considered for further analysis. The application of the ROC curve methodology, together with its limitation and possible extensions, is illustrated by a simple simulation study that compares three score-based DIF methods (Mantel-Haenszel, standardization and Delta plot).

DIF 4b: Evaluation of Multilevel Approaches for Detecting DIF in Polytomous Items

Graham G. Rifenbark, University of Connecticut; H. Jane Rogers, University of Connecticut

Investigations of differential item functioning (DIF) continue to be of interest and importance to measurement practitioners. Item Response Theory (IRT) provides a strong theoretical framework for DIF detection procedures; however, the multilevel linear modeling (MLM) framework provides an alternative that has several advantages. Kamata's (2001) multi-level formulation of IRT models (i.e., HGLM) provides an optimal solution for not only detecting DIF items, but also for explaining DIF via contextual effects (Williams & Beretvas, 2006). Vaughn (2006) extended the HGLM approach for DIF detection to polytomous items and clustered response data. One of the most widely used methods for DIF detection in practice is the Mantel-Haenszel statistic (MH; Holland & Thayer, 1988). French & Finch (2013) investigated a multilevel extension of the MH statistic for clustered response data. They investigated adjustments include matching with respect to a MLM predicted matching score (Pommerich, 1995) and an augmented MH statistic that takes into account clustering (Begg, 1999). Both of these procedures can be extended to the polytomous case (French & Finch, 2013). The purpose of this study is to compare Type I error rates and power for the adjusted MH statistics and the HGLM approach, estimated using a Hamiltonian Monte Carlo routine available in Stan (Stan Development Team, 2016), in the context of polytomous response data. The relative effectiveness of these procedures for polytomous response data has not previously been studied. A simulation study is in progress in which number of clusters, sample size per cluster, and DIF effect size are manipulated.

DIF 4c: Improving Wald test DIF detection in the CDM framework

Yu Bai, Teachers College, Columbia University; Yan Sun, Rutgers, the State University of New Jersey; Charles Iaconangelo, Rutgers, the State University of New Jersey; Jimmy de la Torre, The University of Hong Kong

The Wald test has been adopted in both IRT (Lord, 1980) and CDM frameworks (Hou, de la Torre, & Nandakumar, 2014) for testing DIF. However, its Type I error and power may vary depending on how well the variance-covariance matrix is estimated. In the CDM framework, it is not clear whether the block diagonal variance-covariance matrix introduced in de la Torre (2009) leads to accurate results. In the current study, we compare the performance of the Wald test based on the block diagonal, full analytical variance-covariance matrix, and full variance-covariance matrix estimated via the supplemented EM (SEM) algorithm (Meng & Rubin, 1991). The DINA model and the A-CDM (de la Torre, 2011) are implemented in the Wald test for DIF.

A simulation study is conducted, where the data are generated under the DINA model and A-CDM model using a fixed number of attributes (K = 5) and a fixed test length (J = 30). Two sample sizes (N = 500 and 1000) are considered. Slip and guessing parameters are manipulated to cover high (g=s=0.1), medium (g=s=0.2) and low (g=s=0.3) quality items. Three DIF sizes: none (|Δs| and |Δg| = 0), small (|Δs| and |Δg| = .05) and large (|Δs| and |Δg| = .1), as well as two DIF types (uniform and nonuniform) are considered. Preliminary results showed that, under the DINA model, using the standard errors from the full analytical variance-covariance matrix led to little improvement in the Type I error and power of the Wald test.

DIF 4d: Exploring the Utility of Composite Group Approach to DIF Analysis

Yuxi Qiu, University of Florida; Anne Corinne Huggins-Manley, University of Florida

Conventionally, analyses of differential item functioning (DIF) are conducted by comparing the probability of correct response between majority versus minority. However, defining majority as the reference seems to be less compliant with the definition of fairness as defined by the Standards for Educational and Psychological Testing (APA, AERA, & NCME, 2014). This research is proposed to evaluate the utility of the composite group approach, which has been demonstrated to have several advantages pertaining to detection and interpretation of DIF (Sari & Huggins, 2014). In this simulation study, data are generated based on the 76-item GRE Verbal Reasoning test (Schaeffer et al, 1993). Mean ability difference, relative group sizes, and number of DIF items are manipulated to discover how the composite group approach to DIF extends to situations of unbalanced group sizes and varied numbers of items with non-negligible DIF. Raju's unsigned area method is used for DIF analysis (Raju, 1988). Findings of this research provide guidelines for practitioners and further stimulate a better practice toward evaluating fairness at the item level.

Equating- EQUATE 1: 1:30 PM - 3:00 PM

EQUATE 1a: Detecting Unstable Items on Pre-Equated Tests via Post-Equating Checks

Keyin Wang, Michigan State University

When item response theory (IRT) is used, the terms “pre-equating” and “post-equating” suggest where the equating fits into the assessment cycle. Items on pre-equated tests are calibrated prior to test administration, usually a previous field-test. Thus, for pre-equated tests, IRT-based equated scores are calculated prior to test administration. For post-equated tests, IRT-based equated scores are determined by a calibration occurring after the test is administered. Pre-equated tests are increasingly common in state assessments because of policy and practical considerations. However, the convenience of pre-equating requires additional scrutiny. On pre-equated tests, any kind of item instability, such as item parameter drift (IPD), introduces increased equating error, negatively impacting scoring and performance-level classification (Clark & Kingston, 2003). Although there is no operational post-equating to correct for such problems, post-hoc checking can be done to prevent the propagation of equating error to future tests. Unfortunately, little previous research has directly addressed this issue, resulting in incomplete guidance for conducting such “post-equated checks.” This study proposes possible procedures and begins to evaluate them. Two types of item instability are simulated: 1) IPD of either 4 or 6 items with magnitudes of .2 and .4 on true b parameter; and 2) low-ability IPD for which low-ability students have a higher probability of correct response on the operational test than on the field-test. A series of criteria including the IRT bb-plot method (Sukin, 2010), the delta-plot method (Michaelides, 2003) and the weighted mean absolute difference (WMAD) will be evaluated for their ability to detect unstable items.

EQUATE 1b: Kernel Equating for Non-Equivalent Groups uUing Propensity Scores

Gabriel Wallin, Umeå University, Sweden; Marie Wiberg, Umea University, Sweden

In observed score test equating the objective is to find a function that transforms the test scores of one test form to the scale of another test form. If the test groups are non-equivalent and no anchor test is available, it has been suggested to use covariates which correlate with the test scores to adjust for group differences in ability. This is known as the non-equivalent groups with covariates design. A problem is that already for a relative small dimension of the covariate vector, the possible combinations for the covariates grow rapidly, yielding few, if any, observations for several of the combinations. In this paper we instead suggest to use a scalar function of the covariates known as the propensity score. The propensity score is incorporated within the kernel equating framework and the overall aim was to investigate whether the propensity score can replace raw versions of the covariates. We evaluate our equating estimator in a simulation study and by using real data from an admission test. The resulting equatings are compared using different model specifications of the propensity score and the equatings are also compared with the results from using an equivalent groups design and the non-equivalent groups with anchor test design. The obtained results are promising and shows that using propensity scores in kernel equating is a flexible alternative to adjust for differences in group ability.

EQUATE 1c: IRT Observed Score Equating with the NEC Design

Valentina Sansivieri, University of Bologna; Marie Wiberg, Umea University, Sweden

To be able to compare different test scores from different test forms we use test score equating. Although it is preferable to use a non-equivalent with anchor test (NEAT) design it might be impossible to administrate an anchor test due to test security or other reasons but we still know that the groups are non-equivalent which rules out the equivalent groups (EG) design. A possibility is then to use non-equivalent groups with covariates (NEC) design (Wiberg & Bränberg, 2015). The overall aim was to propose the use of Item Response Theory (IRT) with a NEC design. We start from the Mixed-Measurement IRT with covariates model (Tay, Newman & Vermunt, 2011, 2015). This is used to model associations of latent classes and covariates with external observed characteristics. This model is used within the IRT observed score framework. The proposed test equating method is empirically examined using simulations and empirical data from a large scale assessment of eight graders and a college admission test. The obtained results are compared with IRT observed score equating methods within the EG and the NEAT designs using both simulations and empirical data. The results from the simulations shows that the standard errors of the equating are lower when covariates are included in the IRT model than if they are excluded. The two empirical data sets illustrate the advantages of using this equating method in practice.

Invited Speakers: Ellen Hanmaker, Ed Merkle: 3:15 PM - 4:00 PM

Invited Speakers: Ellen Hamaker: At the frontiers of dynamic multilevel modeling

Ellen Hanmaker, Methodology and Statistics, Utrecht University

Chair: Jee Seon Kim

Due to technological developments (e.g., smartphones), there is an enormous increase in studies based on daily diaries, ecological momentary assessments, ambulatory assessments, and experience sampling methods. The intensive longitudinal data resulting from these studies provide us with the unique opportunity to investigate the dynamics of psychological processes as they are unfolding over time. This has led to a new class of statistical models, which we may call dynamic multilevel modeling: such models are based on using time series models at level 1 to capture the dynamics of the within-person process, while at level 2 we allow for between-person differences in the parameters of these processes. In this presentation I will give an introduction into this new area, including consideration of applications of univariate and multivariate multilevel autoregressive models (which includes dynamical networks). In addition, I discuss some of the specific challenges associated with this kind of statistical modeling.

Bayesian SEM: Some Computational Advances and Potential Pitfalls

Ed Merkle, University of Missouri

Chair: Sophia Rabe-Hesketh

The literature on structural equation models has been slow to incorporate the Markov chain Monte Carlo advances of the 1980s and 1990s, as compared to the literature on other classes of statistical models. While a variety of methods exist (most notably, the methods available in Mplus) for specific types of SEMs with specific types of prior distributions, it remains difficult to estimate many models using general (non-conjugate) prior distributions and to utilize modern Bayesian metrics. In this talk, I will discuss and illustrate some approaches to Bayesian SEM estimation that allow us to harness open source MCMC samplers (i.e., JAGS) and software (i.e., lavaan) so that it is relatively easy to specify, estimate, and extend Bayesian structural equation models. I will also discuss some difficulties associated with Bayesian SEM that have not received much attention in the literature.

Symposium 6: Methodological Advances for Computer-based International Assessments: 4:10 PM - 5:40 PM

Symposium 6a: Modeling Mode Effects in International Large Scale Assessments

Matthias von Davier, Educational Testing Service; Lale Khorramdel, Educational Testing Service

Computer-based assessments provide more information about test-takers response behavior (e.g. process and timing data) leading to more efficient and accurate measurements. They also offer the opportunity to assess new constructs or new aspects and facets of constructs using interactive item types. However, the change from a paper-based to a computer-based mode in international large-scale assessments also challenges the measurement of trend over time. Mode effects might exist in the form of differential item functioning (DIF) observable on (at least some) of the trend items when comparing equivalent groups across paper and computer versions of the assessment. Our study presents an approach on how to test for possible mode effects, and how to best model the effect should it be present. The use of graphical model tests based on equivalent groups designs and different item response theory (IRT) based mode effect models are introduced. IRT model extensions using different mode effect parameters provide information about whether the mode effect is best described by an overall difference between assessment modes, whether it is a person specific effect that may have an impact differentially on different groups, or whether it is an effect that is impacting some subset of tasks. Applied to data coming from the Program for International Student Assessment (PISA), overall model fit will be reported and it will be described how the best fitting model can be used to adjust any mode effect. The mode impact is evaluated compared to gender and random split of schools within each country.

Symposium 6b: Latent and Undirected Graphical Model for Multivariate Binary Data

Yunxiao Chen, Columbia University; Jingchen Liu,

We consider the modeling, inference, and computation for analyzing multivariate binary data. The proposed model combines two popular approaches to modeling multivariate categorical data. In our model, the dependence is mostly induced by a latent vector that usually admits a low dimension and has practical interpretations in different contexts. Such models are very popular in many disciplines. In this study, we consider the context of cognitive assessment, in which the observed variables are responses to items and the latent vectors are interpreted as subjects' underlying attributes. Human cognitive process is typically very complicated and thus is hardly completely driven by just a few factors. Then, a low-dimensional latent factor model is often insufficient to capture all the variations of the data. We propose a model based on the multi-dimensional item response theory and further include a graphical structure capturing the remainder dependence additional to the latent structure. The dependence among the responses is induced by both the latent vector and the conditional graph. To distinguish these two sources and to ensure model identifiability, we impose conditions on the latent and the graphical structures. Corresponding estimation and computational methods are developed.

Symposium 6c: Modeling the Complex Relation Between Ability and Response Time

Hyo Jeong Shin, ETS; Matthias von Davier, Educational Testing Service; Steffi Pohl, Freie University at Berlin

The introduction of computer-based international assessments has made response time data more accessible. The use of timing information in addition to item responses in IRT models can improve ability estimation, item calibration, and the diagnosis of aberrant response behaviors. Recent modeling approaches treat response time as a unidimenional entity and estimate the correlation between the ability and the response time dimension (e.g., van der Linden, 2007). However, the relation between ability and response time may not be that simple or uniform. A couple of recent studies addressed the interaction between ability and response time beyond the linear relation (e.g., De Boeck & Partchev, 2012). In our study, we address the complex nature of response time data and illustrate a new approach using cognitive data from the Programme for International Student Assessment (PISA) collected in 2015. First, we examine the dimensionality of response time using principal component analysis. Second, we show that the relationship between ability and response time can be varied depending on students' ability level, item difficulty level, and aberrant response behavior. We define the response time dimension as “efficient use of time”, and propose a new approach that takes item features and response tendencies into account. In the proposed approach, different parameters depending on the item features are allowed, and missingness coming from omitted responses is incorporated. Bayesian estimation with Markov chain Monte Carlo (MCMC) computation is used for empirical illustration.

Symposium 6d: Feature Generation and Selection Using Process Data from Problem-Solving Items

Qiwei He, Educational Testing Service; Zhuangzhuang Han, Teachers College, Columbia University; Matthias von Davier, Educational Testing Service

Technical advances in computer-based testing have made greater efficiency possible and increased the effectiveness of assessment, including the capability to administer dynamic and interactive problems, engage students' interest more fully, and capture more information about the problem-solving process. A variety of timing and process data such as action sequences can be recorded in log files accompanying test performance data when students respond to items. What we can learn and how we can extract informative features from the process data are essential questions to be answered. The present study analyzes process data collected from a released problem-solving item Climate Control (N = 30,224 students from 42 countries) in PISA 2012 with a focus on generating and selecting informative features that are highly associated with students' performance. The purpose of this study is twofold: first, using process data to generate predictive features in a simulation-based environment; and secondly, to identify robust features that are associated with success or failure on the task. We generated features via two approaches: disassembling action sequences into mini-sequences with n-grams, and creating features that reflect solving strategies, goal directed behaviors, and latency information. Two machine learning methods, random forest and gradient boosting trees were used to select discriminative features for different performance groups. For the sample in hand, in an examination of process data, 15 features were selected out of 78 with a promising prediction accuracy of over 84%. Partial dependency among some selected features is also discussed in the paper.

Psychometrika Anniversary Session 2: 4:10 PM - 5:40 PM

Psychometrika Anniversary Session 2a

Larry Hubert, University of Illinois at Champaign-Urbana

Lawrence Hubert on Johnson (1967)

Psychometrika Anniversary Session 2b

Jacqueline Meulman, Mathematical and Applied Statistics Group

Jacqueline Meulman on Kruskal (1964)

Psychometrika Anniversary Session 2c

Charles Lewis, Fordham University

Charles Lewis on Tucker and Lewis (1973)

Psychometrika Anniversary Session 2d

Willem Heiser, Leiden University

Willem Heiser on Kaiser (1958) by Henk Kiers; online comments on Kaister (1974) by Anonymous

Item Response Theory- IRT 7: 4:10 PM - 5:40 PM

IRT 7a: Truncated Logistic Function Item Response Theory Model

Jaehwa Choi, The George Washington University

New item response theory models supported on [0, 1] bounded interval examinee proficiency are proposed. The proposed models are based on a truncated logistic function, and the rationales and analytic frameworks of the models are articulated parallel to the traditional logistic function item response theory models. The proposed models are established at the expense of some increase in the link function complexity, but greatly boost the interpretability and utility of the person and item difficulty parameters.

For applied researchers, this means a) communicating a clearer characterization and conceptualization of person parameter on the [0, 1] continuum which is more theoretically relevant for some latent attributes (e.g., mastery of a knowledge or skill); b) better interpretation of person parameter lower bound 0 (e.g., absence and/or possible minimum of a skill) and the upper bound 1 (e.g., full mastery and/or possible maximum of a skill); and c) easier interpretation of the pseudo-guessing parameter as the chance of getting the correct answer on an item for a subject with 0 latent trait without introducing possibly ambiguous asymptote concept.

The parameter recovery of the proposed model is evaluated through simulated data. A popular empirical example, five dichotomously scored items on the Law School Admissions Test, is also analyzed to illustrate the proposed approach compared to the traditional models using a Markov chain Monte Carlo (MCMC) estimator. Having illustrated how new model parameters can be straightforwardly estimated via MCMC estimation method without great difficulty over the traditional model cases, the other contribution becomes possible.

IRT 7b: Extending IRT to Tests Containing Matching Items

Matthew D. Zeigenfuse, Universität Zürich; William H. Batchelder, University of California-Irvine; Mark Styvers, University of California, Irvine

This talk presents an IRT approach to modeling behavior on tests containing one or more sets of matching items. For each set of matching items, test takers must associate with each item one element of a set of response alternatives in such a way that no response alternative is offered as the response to more than one test item. Since response alternatives cannot be given as responses to multiple items, responses within a matching set are locally dependent. The model deals with this dependence by introducing a binary matrix of latent "knowledge" parameters whose value depends stochastically on ability through the 2PL model and modeling test taker's response behavior given this matrix. For tests containing multiple sets of matching items or additional non-matching items, the model also incorporates a testlet component to account for residual correlation among responses within matching sets. The ability of the model to account for matching data is demonstrated through a simulation study comparing the matching test 2PL model and testlet 3PL model on test comprised of multiple sets of matching items. This study finds that matching test 2PL model is accurate and does a better job of accounting for test taking behavior than the testlet 3PL, indicating that IRT approaches can be extended to tests with matching items so long as the dependence structure these items is accounted for by the IRT model.

IRT 7d: Linking Polytomous Response Model Parameters with Optimal Linking Design Application

Michelle D. Barrett, Pacific Metrics Corporation; Wim van der Linden, Pacific Metrics Corporation

Response model parameters must be adjusted for the impact of the identifiability restrictions imposed on them during calibration if they are to be compared across calibration studies. Although this holds true for polytomous as well as dichotomous response model parameters, closed-form expressions for asymptotic standard errors (ASE) of the linking parameters for polytomous models have yet to be presented in item response theory literature. In this session, we present derivations of the ASE of linking parameters for items calibrated with common polytomous response models using a multivariate delta method. We then illustrate the use of the ASE for each common item in estimation of the overall linking parameters as precision-weighted averages, a method that favors the items with the least amount of error in their parameter estimates. The precision-weighted average estimator also allows us to make explicit the relationship between the number of response categories of the common items in a linking study and their contribution to the estimation error in the linking function. For the generalized partial credit model, we present an analysis suggesting an ASE for the slope parameter in the linking function that hardly changes with the number of response categories of an item, while the ASE for the intercept parameter increases with the number of categories. We conclude with a simulation study and an empirical linking design study, in which we select an optimal set of common items from an item pool with items of varying numbers of response categories, both confirming the analytic results.

IRT 7c: A Graded Response Model for Addressing Extreme Category Avoidance
Wes Bonifay, University of Missouri-Columbia; Anthony Rodriguez, University of California, Los Angeles

Certain test items may entice respondents to avoid endorsing the extreme options of an ordered response scale (e.g., the "never" or "always" categories of Likert-type scales). Whatever the cause of this behavior, extreme category avoidance may result in inaccurate estimates of the respondent's location along the latent trait scale. We address this by introducing an extension of the Graded Response Model (GRM: Samejima, 1969). The GRM estimates the probability of responding in a given category by reorganizing the polytomous response options as a succession of dichotomies (e.g., 0 vs. {1,2,3}; {0,1} vs. {2,3}; {0,1,2} vs. 3) and then fitting a 2-parameter logistic (or normal ogive) model to each dichotomy. To allow for extreme category avoidance, our approach fits a 3-Parameter Logistic (3PL; Birnbaum, 1968) function to the dichotomies representing the end points of the response scale. Specifically, the proposed 3-Parameter Logistic GRM (3PLGRM) utilizes a 3PL model for the 0 vs. {> 0} dichotomy and an inverted 3PL model to the m vs. {< m} dichotomy, where m is the highest response category. The 3PLGRM thereby accounts for the possibility of a respondent selecting option 1 instead of 0 (or m – 1 instead of m) even though such a response does not reflect his/her true location along the latent trait continuum. In addition to discussing 3PLGRM parameter estimation and computation of item information, we also conduct a sensitivity analysis aimed at uncovering the degree of extreme category avoidance that demands the use of the 3PLGRM rather than the traditional GRM.

Diagnostic Classification Model- DCM 4: 4:10 PM - 5:40 PM

DCM 4a: Assessing Change over Time in a General Diagnostic Classification Model
Matthew J. Madison, University of Georgia; Laine Bradshaw, University of Georgia

One of the most common assessment research designs is the pre-test/post-test design in which examinees are administered an assessment before instruction, and then take the same, or a parallel form of the assessment after instruction. In this type of study, the primary objective is to measure growth in examinees, individually and collectively. In an item response theory (IRT) framework, longitudinal IRT models can be used to assess change in examinee ability over time. In a diagnostic classification model (DCM) framework, assessing growth translates to measuring change in attribute mastery over time. The focus of this study is to develop and examine new methodology to accommodate this type of longitudinal data in a general DCM. More specifically, this study combines latent transition analysis (LTA) with the log-linear cognitive diagnosis model (LCDM) to model transition and change in attribute mastery in a pre-test/post-test designed study. Previous work has combined LTA with a constrained DCM. Therefore, the proposed model, referred to as the LTA-LCDM, is a generalization and more applicable extension of previous work. Preliminary results show that the LTA-LCDM has good item parameter recovery, accurate and reliable classifications, appropriate type I error and strong power in detecting growth. Through an empirical analysis, we show how the LTA-LCDM can be used to provide valuable diagnostic feedback on examinee mastery transition and growth.

DCM 4b: Measuring Cognitive Processing Capabilities in Solving Mathematical Problems
Susan Embretson, Georgia Institute of Technology

Diagnosing the skills possessed by examinees in solving items, particularly mathematical items, have been assessed in many settings with diagnostic item response models (Rupp, Templin & Henson, 2010; von Davier, 2008). However, these skills are typically not assessed in the context of a cognitive model of processing complexity. Studies have shown that the levels and sources of cognitive complexity predict item difficulty (e.g., Embretson & Daniel, 2008) and, further, that items can be selected or designed for difficulty in different sources of cognitive complexity. Although these results are relevant to the response processes aspect of construct validity, potential impact on the other aspects of validity was not addressed. That is, the modeling procedure did not include multidimensional measurement of individual differences in processing capabilities. In the current study, the multicomponent latent trait model for diagnosis (MLTM-D, Embretson & Yang, 2013) was applied to measure cognitive processing capabilities in the translation, integration and solution execution stages of processing the items. Individual differences in the patterns of processing capabilities were significantly related to examinee background variables, thus indicating potential impact on the consequential aspect of validity. Implications of the findings for item design and test development will be discussed.

DCM 4c: Possibilities of the Rule Space Method in Large Scale Assessments

Thorben Huelmann, University of Zurich

The Rule Space Method (RSM) is a diagnostic competency model. Within this model, students are described by at least two dimensions: the ability of the student and the unusualness of the answering pattern. An answering pattern therefore is considered unusual, if a student fails at easy items, but succeeds at difficult items. Unusual answering patterns can be created by various reasons. For example, difficult items can appear as an easy item to a cheating student. The cheating would lead to some turmoil in the order of difficulty of the items and with this leading to unusual answering patterns. If a Q-Matrix exist, which links competencies to items, ideal answering patterns can be created. With the help of these ideal answering patterns, students can be classified by their competency. The classification is not forced upon the data, so that single students can remain unclassified. In my presentation, I will first illustrate how the RSM works. Then, I will show an application of the RSM. Therefore, I will be using the TIMSS 2007 dataset. I am going to close the presentation with an outlook of further possible applications of the RSM and its implementation in R.

DCM 4d: Classification Accuracy Comparison Between Measurement Decision Theory and IRT

Yating Zheng, University of Maryland, USA; Hong Jiao, University of Maryland

Measurement decision theory, which derives from Bayesian theorem, is an attractive alternative to IRT in making classification decisions. Its key idea is to get a best estimate of an examinee's mastery state based on the examinee's item responses, item parameters and prior population classification proportions. Previous research indicates that measurement decision theory requires fewer items and smaller sample sizes to achieve the same level of classification accuracy than the IRT models. However, these studies primarily focused on tests with either dichotomously or polytomously scored items. This study explores the application of measurement decision theory in mixed format tests which contain both dichotomous and polytomous items. Comparison is made between classification accuracy from the concurrent use of the 2PL IRT model and the generalized partial credit model and that from measurement decision theory. Several factors --- test length, sample size, and the proportion of dichotomous and polytomous items --- are manipulated to simulate different study conditions . It is expected that measurement decision theory provides higher classification accuracy than the IRT models. Further, inferential statistical analysis using multiple regression is conducted to check significance of the manipulated factors on classification accuracy.

Applications- APP 3: 4:10 PM - 5:40 PM

APP 3a: The Effects of Test Preparation on Students' State Test Performance

Yao Xiong, Penn State University; Hongli Li, Georgia State University

As a result of the increasing emphasis on the accountability policy and large-scale assessment, teachers are incentivized to adopt more test preparation activities in their classes. Due to the controversial nature of test preparation and the lack of high-quality data, systematic examinations have been rarely conducted on test preparation in K-12 settings. Drawing on the Measure of effective teaching (MET) dataset, the present study examines to what extent students are involved in test preparation activities and the effects of test preparation on students' state test performance. The MET is the largest study of classroom teaching ever conducted in the United States (Bill and Melinda Gates Foundation, 2012). The Year 2 dataset was used for the analysis, in which 1,375 4th-9th grade teachers from 278 schools participated in the study. Based on students' responses to a perception survey, we found that test preparation practice was rather frequent and prevalent. Students had slightly more test preparation activities in Math than in English Language Arts (ELA), and students at lower grade levels reported having more test preparation activities. Furthermore, students with lower scores in Year 1 tended to be involved in more test preparation activities in Year 2. Different test preparation practices seemed to have different effects on students' state test performance, though the effects were generally small in a practical sense. The results of this study provide important knowledge on test preparation as a result of the current accountability policy in the K-12 settings.

APP 3b: Comparison Decision Accuracy Complex Decision Rules in Higher Education Context
Iris Yocarini, Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, The Netherlands

This study systematically evaluated the decision accuracy of complex decision rules combining multiple tests within different educational settings. Fully compensatory, fully conjunctive, and complex decision rules mixing both aspects were evaluated. In a fully compensatory testing system, students are allowed to compensate a low score on one course with a high score on another course. By contrast, students in a fully conjunctive testing system are required to pass each individual course. When both a minimum grade (conjunctive aspect) and a minimum average level of performance (GPA; compensatory aspect) is required, as is most common in educational decisions, a complex decision rule is in place. Simulations were performed to obtain students' true and observed score distributions and to manipulate several factors relevant to educational settings in practice. The results showed that the decision accuracy depends on the conjunctive (required minimum grade) and compensatory (required GPA) aspects and their combination. Overall, within a complex compensatory decision rule, the sensitivity is higher and specificity lower compared to a conjunctive decision rule. For a conjunctive decision rule, the reverse is true. Which rule is more accurate also depends on the average test reliability, average test correlation, and the number of retakes that are allowed. This comparison highlights the importance of evaluating decision accuracy for high stakes decisions, considering both the specific rule as well as the selected measures.

APP 3c: The Effects of Test Preparation on Students' State Test Performance
Hongli Li, Georgia State University; Yao Xiong, Penn State University

As a result of the increasing emphasis on accountability policy and large-scale assessment, teachers are incentivized to adopt more test preparation activities in their classes. Due to the controversial nature of test preparation and the lack of high-quality data, systematic examinations have been rarely conducted on test preparation in K-12 settings. Drawing on the Measure of effective teaching (MET) dataset, the present study examines to the extent to which students are involved in test preparation activities and the effects of test preparation on students' state test performance. The MET is the largest study of classroom teaching ever conducted in the United States (Bill and Melinda Gates Foundation, 2012). The Year 2 dataset was used for the analysis, in which 1,375 4th-9th grade teachers from 278 schools participated in the study. Based on students' responses to a perception survey, we found that test preparation practices were rather frequent and prevalent. Students had slightly more test preparation activities in Math than in English Language Arts (ELA), and students at lower grade levels reported having more test preparation activities. Furthermore, students with lower scores in Year 1 tended to be involved in more test preparation activities in Year 2. Different test preparation practices seemed to have different effects on students' state test performance, though the effects were generally small in a practical sense. The results of this study provide important knowledge on test preparation as a result of current accountability policies in K-12 settings.

APP 3d: Measuring Student Engagement During Collaboration
Peter Halpin, New York University; Alina A. von Davier, Educational Testing Service, USA; Jiangang Hao, Educational Testing Service; Lei Lui, ETS

This research addresses performance assessments that involve collaboration among students. We begin by decomposing the statistical dependence in a collaborative performance assessment into a part that depends on interactions between students (inter-individual dependence) and an additional part that depends on only the actions of individual students considered in isolation (intra-individual dependence). We then discuss the use of the Hawkes process as a parametric modeling framework that captures these two sources of dependence. In particular, the Hawkes process is useful for inferring whether the actions of one student are associated with increased probability of further actions by his / her partner(s) in the near future. This leads to an intuitive notion of engagement among collaborators. We propose a model-based index that can be used to quantify the level of engagement exhibited by individual team members, and show how this can be aggregated to the team level. We also present results on the standard error of the proposed index, which allows for considerations about how to design tasks such that engagement can measured reliably. The approach is illustrated using a simulation-based task designed for science education, in which pairs of collaborators interact using online chat. We also consider the empirical relationship between chat engagement and task performance, finding that less engaged collaborators were less likely to revise their responses after being given an opportunity to share their work with their partner.

Latent Class Analysis- LCA 1: 4:10 PM - 5:40 PM

LCA 1a: Comparison of DINA Model and Exploratory Latent Class Analysis

Tahsin Oğuz Başokçu, Ege University, Turkey; Duygu Güngör, Izmir University

Deterministic Input Noisy and Gate (DINA) model is a latent class model like most other Cognitive Diagnosis Model (CDM). DINA model is based on relation between item and latent variable (skill, attribute etc.). The model depends on true identification of structures as well as items which are needed to be answered correctly for particular skill or skills. Exploratory latent class analysis can be conducted to find number and structure of the latent classes without any a priori information, however. By using LCA one can assign participants to their most likely classes. This research aims at comparing these two models with a simulation study. We analysed simulation data using both DINA and LCA and detected their similarity in assigning participants having specific skills. Additionally, we investigated LCA results as if item response probabilities pointed out same structure defined by Q-matrix. Our simulation conditions were related to number of items (18, 22 or 26) and number of attributes (2, 3 or 4), sample size was fixed to 2000. As a result, regardless to item numbers when the attribute number was two, LCA pointed out four class solution and .98 of the participants were assigned same classes. Similarly, when the attribute number was three, LCA pointed out eight classes and .88 were assigned same classes. However, when the attribute number increased LCA failed to find same results with DINA.

LCA 1b: Identifiability of Cognitive Diagnostic Models

Gongjun Xu, University of Minnesota

Statistical latent class models are widely used in social and psychological researches, yet it is often difficult to establish the identifiability of the model parameters. In this talk we consider the identifiability issue of a family of restricted latent class models – Cognitive Diagnostic Models – where the restrictions are needed to reflect pre-specified assumptions on the related assessment. We establish the identifiability results in the strict sense and specify which types of items would give the identifiability of the model parameters. The results not only guarantee the validity of many of the popularly used models, but also provide a guideline for the related experimental design.

LCA 1c: Aggregate Constant Ratio Model Clustering

Stephen L. France, Mississippi State University; Sanjoy Ghose, University of Wisconsin-Milwaukee

The aggregate constant ratio model (ACRM) is an aggregation of Luce's choice axiom and assumes a situation where the probability of choosing one item over another is not affected by other items in the choice set. This presentation describes a model for utilizing deviations from ACRM choice data to cluster items based on overall choice similarity. The model uses a maximum-likelihood methodology and is fit using a heuristic local search optimization procedure. A variant of the “gap statistic” is used to help choose the number of clusters. A k-fold cross validation procedure is utilized to test the stability of the clustering solutions. Applications to brand switching behavior and to more general psychological choice model scenarios are given. Prior intuition can be tested against the computationally derived clustering solutions using a visual methodology. A computational package, programmed in R, is described.

Networking Session on Analysis of Intensive Behavioral Data from Device-Enabled Studies: 4:10 PM - 5:40 PM

Networking 1a- Session on Analysis of Intensive Behavioral Data from Device-Enabled Studies
Mariejke E. Timmerman, University of Groningen, Heymans Institute for Psychological Research, Psychometrics and Statistics

The Psychometric Society is sponsoring this networking event around analysis and related psychometrics of data from device-enabled studies . Particularly relevant modeling activities include state space, threshold models, MTMM, as well as the (limited) range of models that deal with parsimony for trend data. The goal of the meeting is to identify interested parties and also inform articulation of a cross-section of the inferential needs that have arisen in mobile behavior measurement applications. We encourage all members with relevant experience to join and converse about the current set of methods in use in light of current and future scholarship around these needs. Possible member benefits include collective application for conference and project funding, as well as the development of formal and informal scholarly networks.

Networking 1b- Session on Analysis of Intensive Behavioral Data from Device-Enabled Studies
Wen Chung Wang, The Hong Kong Institute of Education

The Psychometric Society is sponsoring this networking event around analysis and related psychometrics of data from device-enabled studies . Particularly relevant modeling activities include state space, threshold models, MTMM, as well as the (limited) range of models that deal with parsimony for trend data. The goal of the meeting is to identify interested parties and also inform articulation of a cross-section of the inferential needs that have arisen in mobile behavior measurement applications. We encourage all members with relevant experience to join and converse about the current set of methods in use in light of current and future scholarship around these needs. Possible member benefits include collective application for conference and project funding, as well as the development of formal and informal scholarly networks.

Career Award Speaker: Wim van der Linden- Big Data, Small Step
Wim van der Linden, Pacific Metrics Corporation

Chair: Terry Ackerman

Abstract is not available at this time

Symposium 7: The Development and Use of Noncognitive Assessment in K-12: 9:45 AM - 11:15 AM

Symposium 7a: Overview: Noncognitive assessment in K-12

Patrick Kyllonen, Educational Testing Service

Over the last 10 years there has been a growing awareness of the importance of noncognitive factors in education. Partly this comes from studies that show that the well-documented benefit of educational attainment on subsequent outcomes (e.g., employment, earnings, civic behavior) is only partially due to the cognitive skills acquired in school. Most of the "education effect" is due to noncognitive skills (e.g., Bowles, Gintis, & Osborne, 2001; Heckman, Humphries, & Kautz, 2014). This motivates more attention on noncognitive factors. Several meta-analyses have identified the key noncognitive factors, based on validity studies predicting school outcomes (e.g., Poropat, 2009; 2014) and randomized trials showing benefits for noncognitive skills training (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011). Factors generally fall into the categories of interpersonal and intrapersonal competencies (National Research Council, 2012), social and emotional skills (Durlak, Domitrovich, Weissberg, & Gullotta, 2015), or behaviors, strategies, dispositions, and attitudes (Farrington et al., 2012). There also is a companion literature on assessment methods (Kyllonen, 2016) outlining tradeoffs between psychometric properties and cost, burden, and administrative convenience. Among the most important assessment methods for K-12 education are self-assessments (rating scales and forced-choice methods), others' ratings, and situational judgment tests. We have conducted several surveys with teachers, principals, and school administrators that corroborate the importance of factors such as work ethic, time management, self and emotional regulation, teamwork and collaboration, and others. In this talk I review findings pertaining to constructs and best methods for measuring them as background for the symposium.

Symposium 7b: Development of Character Skills Assessment for Secondary Schools

Jinghua Liu, Secondary School Admission Test Board

Traditionally, the measurement of cognitive skills such as reading, writing and mathematics plays an important role in the K-12 realm. While the cognitive skills will continue to be a vital part of education, educators, researchers and admission professionals have been becoming aware of the importance of skills other than cognitive skills – character skills.

As illustrated by other papers in this symposium, many studies have been conducted and there are preliminary indications that such measures of character skills provide important information about students that relates to readiness and likelihood of succeeding in school (Kyllonen, 2016; Kuncel, 2016). In 2014, after two years' of exploring and studying the character skills landscape, Secondary School of Admission Test Board (SSATB) move forward in developing a character skills assessment tool, which aims to offer insights into an applicant's character attributes, and to provide a holistic view of an applicant: character skills and cognitive skills.

The purpose of this paper is to illustrate the development and psychometric properties of the SSATB's character skills assessment, which includes: constructs to be measured based on research and stakeholders' input; pretest of Likert-type statement; development of forced choice items and situational judgment test items; item analysis; test reliability; and preliminary validity evidence.

Symposium 7c: Examination of the Validity of Measures Assessing Social-Emotional Skills

Katie Buckley, Transforming Education; Sara Bartolino Krachman, Transforming Education

Mounting evidence of the importance of non-tested skills in explaining students' academic success has generated calls to incorporate "social-emotional" skills, or "non-cognitive" skills, into evaluations of educational performance and school accountability systems. Yet we have little evidence on the ability of existing measures of social-emotional skills, most of which are based on student self-reports, to reliably capture differences in student skills across schools. To shed light on this issue, we draw on data on more than 200,000 students in California's CORE Districts, a consortium that is in the process of incorporating measures of social-emotional skills into its school accountability system. In this paper, we provide descriptive evidence on the validity of these measures. First, we consider whether self-report measures of various social-emotional skills (i.e., self-management, social awareness, growth mindset, and self-efficacy) are associated in expected ways with a range of students outcomes (i.e., grades, test scores, absences, and suspensions). Second, for a subset of districts, we compare the predictive power of student self-reports and teacher-reports in explaining student outcomes. Third, we present evidence on the extent to which the anchoring vignettes included in the pilot year worked as expected. Our preliminary findings suggest that these self-report measures can produce valid results of students' social-emotional skills. We do not find clear evidence that students' responses are biased by differences in school climate across schools. However, our results cannot address questions about how the properties of self-report measures would change if stakes are attached.

Symposium 7d: Nominal Response Model Scoring of a Situational Judgment Test

Jiyun Zu, Educational Testing Service; Hongwen Guo, Educational Testing Service; Patrick Kyllonen, Educational Testing Service

A powerful but under-utilized tool in K-12 noncognitive assessment is the situational judgment test (SJT). SJTs describe a situation (e.g., "your team members ignore your suggestion") and present a set of possible responses (e.g., "leave the group," "discuss why you think it's a good idea," "tell the teacher"), and respondents select "the best" and "the worst", or "the most" or "least likely" depending on instructions. SJTs are ideally suited for measuring social and emotional skills, as shown in the Situational Test of Emotional Management for Youths (STEM-Y) designed for middle schoolers (MacCann et all, 2010).

A challenge in scoring SJTs is that there might not be consensus on the correct response, so there may be multiple correct responses in a situation. This can be seen in non-parametric option characteristic curve (OCC) plots using tentative keys such as percentage agreement with the most popular response (Guo et al, 2015). The nominal response model (NRM; Bock, 1972) can model this complexity as each response has its own slope and intercept, and produce partial credit scoring. NRM scores were found to be more reliable than dichotomously scoring methods, or consensus scores (Zu & Kyllonen, 2013). However, as a data-driven approach there is a question about scoring invariance across samples. We analyzed STEM-Y data from two cohorts (N = 13,780) and found that item parameters correlated across years (r = .98), as did ability estimates (r = .99); findings from subgroups based on gender, culture, and SES will also be reported and discussed.

Symposium 7e

Nathan Kuncel, University of Minnesota

Abstract is not available at this time

Psychometrika Anniversary Session 3: Past Editors: 9:45 AM - 11:15 AM

Psychometrika Anniversary Session 3a: Past Editors

Irini Moustaki, Department of Statistics, London School of Economics

Abstract is not available at this time

Psychometrika Anniversary Session 3b: Past Editors

Larry Hubert, University of Illinois at Champaign-Urbana

Abstract is not available at this time

Psychometrika Anniversary Session 3c: Past Editors

Willem Heiser, Leiden University

Abstract is not available at this time

Psychometrika Anniversary Session 3d: Past Editors

Ulf Bockenolt, McGill University

Abstract is not available at this time

Psychometrika Anniversary Session 3e: Past Editors

Brian Junker, Carnegie Mellon University

Abstract is not available at this time

Psychometrika Anniversary Session 3f: Past Editors

Shizuhiko Nishisato, University of Toronto

Abstract is not available at this time

Categorical Data Analysis- CDA 1: 9:45 AM - 11:15 AM

CDA 1a: Logistic Regression with Misclassification in Binary Dependent Variables

Haiyan Liu, University of Notre Dame; Zhiyong Zhang, University of Notre Dame

Misclassification results from the response errors in categorical variables, which leads to the recorded value of a discrete response variable different from its underlying true value. Misclassification can happen in many scenarios. For example, it can happen when the respondent misunderstands a question or simply chooses the wrong answer by accident. It may also happen in a survey when the participants do not want to give a truthful response. For instance, in a study of marijuana use, a participant who has used marijuana might choose not to report it worrying about the potential penalty.

When a binary dependent variable, subject to misclassification, is analyzed by conventional logistic regression, it can result in misleading parameter estimates and statistical inference. Through Monte Carlo simulation studies, we show that there are nonignorable biases in the parameter estimates if misclassification is ignored. To deal with the problems, we introduce logistic regression models including correction parameters. To estimate the model, we develop a Newton-based algorithm, which offers both parameter estimates and standard errors. We also show with the new models, one can not only obtain the underlying association between the independent and dependent variables but also the estimated extent of misclassifications through simulation studies. Finally, an example on the marijuana use is discussed.

CDA 1b: Determining Optimal AUD Diagnostic Criteria Through Heaviness of Consumption

Jordan Stevens, University of Missouri-Columbia; Douglas Steinley, University of Missouri-Columbia; Kenneth Sher, University of Missouri-Columbia

The Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases and Related Health Problems (ICD) have helped clinicians and researchers classify mental disorders for many years. Although standardization of criteria has greatly advanced classification procedures, there are still many that are unconvinced that the DSM or the ICD are valid and reliable diagnostic tools. Many have issues with the false positive rates and the polythetic criterion endorsement imposed for diagnosis (Lasalvia, 2015; Lilienfeld, et al., 2013; Wakefield, 2015). Using suboptimal criterion sets could lead to increased rates of Type 1 (false positive) and Type 2 (false negative) errors in research settings, resulting in misdiagnosis or failure to diagnose in the clinical setting. Therefore, it is imperative to derive more reliable, stable, and concise criterion sets for clinical use to reduce clinical diagnostic burden and increase validity of the diagnosis. Applying optimization techniques with respect to Alcohol Use Disorder (AUD) diagnosis and using known external validators, a more effective and precise system for diagnosis is developed. Using data from the National Epidemiological Survey on Alcohol and Related Conditions (NESARC), heaviness of consumption is used to optimize the number of criterion in the AUD diagnosis set and the threshold for AUD diagnosis. Ten-fold cross-validation techniques are used to determine which solution sets are optimal compared to existing approaches to diagnosis. Optimizing diagnostic criterion sets can generate more reliable rules, leading to less misclassification of diagnosis.

CDA 1c: Classification Consistency and Accuracy for Mixed-Format Tests

Stella Kim, University of Iowa; Won-Chan Lee, University of Iowa

When the main purpose of a test is to evaluate an examinee's status with respect to a predefined criterion, it is important to ensure that the examinee is consistently and accurately classified into the same performance category over replications of the same (or similar) measurement procedure. The current study is designed to explore classification consistency and accuracy for mixed-format tests consisting of a mixture of multiple-choice (MC) and free-response (FR) items. Real data from several large-scale mixed-format exams are used in this study, which have varying characteristics of test reliability, test length, section weights, multidimensionality, and score distributions. In addition, this study compares classification consistency and accuracy estimates based on various classical and IRT procedures. Use of non-integer weights for the MC and FR sections for composite scores is an additional source of complexity, and this paper employs an efficient approach to dealing with the non-integer issue. Also, this study examines the effects of "structural bumpiness" in score distributions resulting from use of non-integer section weights. Finally, the impact of multidimensionality on classification indices is investigated. The results show that classification indices are affected substantially by the cut-score location and test reliability. The relative performance of various estimation procedures is affected by the degree of model fit in terms of fitted observed score distributions, and the degree of multidimensionality in terms of latent correlation between the construct related to item formats. Practical implications for estimating classification consistency and accuracy for mixed-format tests are discussed.

CDA 1d: Classification Accuracy and Consistency Using the Loglinear Model Based Method

YoungKoung Kim, College Board; Tim Moses, College Board; Won-Chan Lee, University of Iowa

Estimating decision indices – classification accuracy and consistency – typically involves the true score and error distributions. Strong true score models have been used to estimate the true score distribution for the procedure of estimating classification accuracy and consistency (Hanson & Brennan, 1990; Livingston & Lewis, 1995). A new estimation method that expands these strong true score models by fitting more moments of tests' true and observed score distributions was recently proposed (Moses & Kim, 2016). In the traditional strong true score models, the four-parameter beta true score distributions are estimated using "method of moments" methods (Hanson, 1990), which can fit the moments only up to the fourth moment and often have problems fitting the fourth moment. With this new method, fitting additional moments of the true score distribution is easily accomplished by incorporating loglinear presmoothing methods (Holland & Thayer, 2000).

The main purposes of the current study are to apply the loglinear model based method to the procedure of classification accuracy and consistency and to compare this new method with traditional strong true score model methods. Data obtained from a large-scale assessment are used to evaluate the classification accuracy and consistency with the cut scores set on the composite scores, which are the summed scale scores. Classification accuracy is quantified based on false-positive and false-negative error rates.

Classification consistency is evaluated using agreement index and Cohen's kappa. Three cut scores that are associated with the 25th, 50th, and 75th percentiles are examined in terms of the magnitudes of these decision indices.

Multilevel/Hierarchical/Mixed Methods- MLM 2: 9:45 AM - 11:15 AM

MLM 2a: Mixtures of IRT, Generalizability Theory and Multilevel Models

Cees Glas, University of TwenteE.A. van der Scheer, University of Twente, the Netherlands; A.J. Visscher, University of Twente, the Netherlands

In educational research, part of the data may consist of observations by multiple raters at different time points collected using itemized observation instruments. Further, the data often have a multilevel structure pertaining to students nested within teachers or classes. The model for analyzing such data can involve the combination of an IRT model for the item responses, a generalizability theory model for different raters at different time points, and a multilevel model to connect outcome variables with predictors.

Three examples are presented. The first one concerns a study of the relationship between differentiated instruction and student achievement. The outcome is student achievement with several level 1 (students) and level 2 (teachers) predictors, including an achievement pretest, and a combined IRT and generalizability theory model for observations on teacher behavior. The second one is the analysis of an intervention study where teacher behavior is assessed by different raters at three time points before and after the intervention. The third one is also an intervention study, but here the observations are made by the students nested within teachers.

An encompassing model for analyzing these data is presented and various Bayesian approaches for parameter estimation are studied. We compare OpenBugs analyses with analyses using dedicated software based on a data-augmented Gibbs sampler. Further, we compare an approach where the complete model is estimated concurrently with a procedure consisting of two steps where the item-parameters of the IRT model are estimated first while the structural latent regression models are estimated in a second step.

MLM 2b: Approximating Level-1 and Level-2 Heteroscedasticity with Nonparametric Multilevel Mixture Models

Jason D. Rights, M.S., Vanderbilt University Quantitative Methods Program; Sonya K. Sterba, Vanderbilt University Quantitative Methods Program

Multilevel data structures are common in the social sciences. Often, such nested data are analyzed with multilevel models (MLMs) wherein between-cluster variability in intercepts and/or slopes is modeled by continuously-distributed random effects. As an alternative approach, it is known that nonparametric multilevel regression mixture modeling (NPMM) can approximate continuous random effects through discrete latent class variation. However, it has not been recognized or shown analytically that NPMM with classes at level-1 and level-2 can also serve as a nonparametric approximation of MLM heteroscedastic residual variance at level-1 and/or level-2. Given that level-1 and level-2 residual heteroscedasticity in MLM is underinvestigated (according to Goldstein, 2011), unfamiliar to many MLM researchers (according to Snijders & Bosker, 2012), and, moreover, cannot be accommodated in some MLM software, NPMM has practical utility as a flexible, exploratory tool to highlight, diagnose, or investigate such heteroscedasticity (Rights & Sterba, resubmitted). In this talk, we show analytically how NPMM approximates level-1 and level-2 residual heteroscedasticity in MLM and illustrate these relationships with simulated graphical demonstrations. Additionally, we demonstrate an R function, NPMM.approximation, that we developed to allow researchers to visualize and implement NPMM as a nonparametric approximation of MLM residual heteroscedasticity. An empirical example is presented that illustrates practical benefits of these approximations. Extensions involving using NPMM to inferentially test for such heteroscedasticity are discussed.

MLM 2c: Introduction to Relational SEM and an Efficient Computational Strategy

Joshua Pritikin, University of Virginia

We introduce relational SEM, an adaptation of structural equation modeling to relational databases. Relational SEM is a superset of the mixed model and multilevel SEM. In addition, we introduce Rampart, a new computational strategy for frequently encountered relational SEM models. Rampart is inspired by the fact that the multivariate normal density is transparent to orthogonal rotation. Well suited to big data, Rampart becomes more effective as the size of the data set increases. When data are strictly nested then there are usually fewer variables in the upper level connected to many more variables in the lower levels. A regression from teacher skill to student performance has this characteristic. In such a model, under typical conditions, a rotation can be applied to eliminate all but one of the links from teacher to student with a corresponding rotation applied to the observations. This transformation leaves the likelihood function unchanged, but offers a major benefit: dramatically increased independence in the model implied covariance matrix. Performance on a variety of models is compared with and without Rampart. Limitations of Rampart are discussed. Rampart requires strictly nested structure and identical sub-models. Rampart can be applied locally to the part of a model that meets these criteria leaving the remaining parts of the model untouched. R source code is included for all models discussed. Rampart is implemented in OpenMx. OpenMx is free and open software that runs on all major operating systems.

MLM 2d: Generalized Linear Mixed-Effects Regression (glimmer) Trees: Model-Based Recursive Partitioning of Multilevel Data

Marjolein Fokkema, Leiden University; Achim Zeileis, University of Innsbruck; Niels Smits, University of Amsterdam; Torsten Hothorn, Universität Innsbruck; Henk Kelderman, Universität Zürich

Model-based recursive partitioning is a useful tool for subgroup detection in, for example, GLMs. The rationale behind recursive partitioning is that an overall parametric model may not fit the data well. When additional covariates are available, it may be possible to partition the dataset with respect to these covariates, and to find better fitting models in each cell of the partition. Each cell of the partition represents a subgroup, with different values for one or more model parameters. However, datasets that are used for the detection of such subgroups may often have a multilevel structure. For example, observations within a dataset may be nested within schools, persons (in longitudinal datasets), hospitals, or research centers. Therefore, we developed glmertree: a tree-based algorithm allowing for the detection of subgroups with different values for one or more parameters of a GLM, as well as estimation of random effects. In this presentation, I will introduce the glmertree algorithm and its R implementation, and present some findings on the performance of glmertree in real and simulated datasets.

Missing Data- MIS 1: 9:45 AM - 11:15 AM

MIS 1a: Combination Rules for Multivariate Analysis of Variance in Multiple Imputation

Joost Van Ginkel, Leiden University; Niels Smits, University of Amsterdam

Multivariate analysis of variance (MANOVA) is a widely used statistical technique within social and behavioral sciences. It extends univariate analysis of variance to a situation with several outcome variables rather than one outcome variable. Like any other statistical technique, the results of MANOVA may be influenced by missing data. Multiple imputation is a well-established procedure to deal with missing data. In multiple imputation, missing data are estimated multiple times, resulting in multiple complete versions of the incomplete dataset. Each of these datasets are then analyzed by the statistical analysis of interest, and the results are pooled into one final statistical test or parameter estimate. However, to date, no rules have been defined for combining the F tests of the multiply imputed datasets in MANOVA. In this presentation combination rules for combining F tests of MANOVA from multiply imputed data sets are proposed, using already existing pooling techniques for multiple imputation.

MIS 1b: Handling Item Non-Response in Structural Equation Modeling with Ordinal Variables

Myrsini Katsikatsou, London School of Economics; Irini Moustaki, Department of Statistics, London School of Economics

Within the framework of Structural Equation Modelling for ordinal variables, the three-stage least squares (3S-LS) and the Pairwise likelihood (PL) estimation methods yield biased estimators, in general, if they ignore missing values which are missing at random (MAR). We develop three different versions of the PL method to deal with MAR missingness: the complete-pair PL (CP-PL), the available-case PL (AC-PL), and the doubly-robust PL (DR-PL). The first two draw on the approach of full information maximum likelihood with missing data, while DR-PL draws on the theory of doubly-robust estimators. The performance of these three PL estimators is studied with respect to the bias and MSE of parameter estimates and standard errors as well as to the easiness of application. Also, their performance is compared to that of the multiple-imputation approach developed under the 3S-LS approach.

MIS 1c: Sequential Cluster Sampling for International Studies

Sewon Kim, Michigan State University; Mark Reckase, Michigan State University; Unhee Ju, Michigan State University

The work reported here was stimulated by the needs of an international study into the ways that new teachers are integrated into the educational system called The First Five Years of Mathematics Teaching (FIRSTMATH) Study. In the development of the research plan for this study it was discovered that new teachers make up a relatively small proportion of the teaching force, they are not uniformly distributed over schools, the population changes frequently, it may be difficult to gain participation, and sometimes the information about when a person started teaching are not readily available. These characteristics of the target population for this study suggest a particular class of sampling problems where the target entity is rare, the distribution over larger units is not known, and the cost of obtaining information about the target entities is fairly high. This led to the development of a sampling process called Sequential Cluster Sampling (SCS), a special case of a more general process called Stratified Sequential Adaptive Cluster Sampling (SSACS). This paper describes this sampling process and reports an evaluation of it compared to simple random and stratified random sampling. Overall, the sampling methodology is shown to be a useful alternative for some types of international studies that have the challenge of recruiting relatively rare individuals to participate.

Symposium 8: Advanced Statistical Power Analysis in Real Scenarios: 9:45 AM - 11:15 AM

Symposium 8a: Power Analysis for T-Test with Non-Normal Data and Unequal Variance

Han Du, University of Notre Dame

A t-test is a statistical hypothesis test in which the test statistic usually follows a Student's t distribution if the null hypothesis is true and follows a non-central t distribution if the alternative hypothesis is true. A t-test often assumes that data are normally distributed while real data typically deviate from normal. In the way of combining variances, different methods for pooling the variances are available especially when the two samples have different population variances. Since the previous methods just approximate the true distribution, different approximation approaches may yield different power values, which leaves a decision for researchers to make. All these pose challenges on statistical power analysis for t-test. We develop a general method to conduct power analysis for t-test through Monte Carlo simulation. The method can flexibly take into account non-normality and unequal variances in two-sample studies. Software is developed to carry out the Monte Carlo based power analysis. Examples are used to illustrate the method and software.

Symposium 8b: Statistical Power Analysis for Multilevel Modeling

Miao Yang, University of Notre Dame

Statistical power analysis plays an important role in research designs. The techniques of power analysis for single-level models under standard assumptions have been well developed and implemented in popular statistical software (e.g., SAS PROC POWER). In educational and psychological studies, however, data often exhibit nested structure, for example, students nested within schools, patients nested within clinics, etc. Ignoring the nested structure will lead to significance tests with inflated type I error rates and misleading power. Multilevel models have been developed to account for the structure of nested data. Several studies have investigated the power for multilevel models by utilizing z-tests or chi-square difference tests. However, few studies take missing data into account. In this study, existing procedures of power analysis for multilevel models are reviewed. A simulation-based approach is developed to estimate the power for multilevel models when missing data exist. The power estimation is implemented in an online software called Webpower. The application of Webpower is illustrated through several examples.

Symposium 8c: Statistical Power for ANOVA with Binary and Count Data

Yujiao Mai, University of Notre Dame

ANOVA is widely used to analyze experimental data for hypothesis testing, while binary and count data are common in data collection. Current power analysis for ANOVA generally assumes that the outcome data are normally distributed. Few studies have discussed how to conduct statistical power analysis for ANOVA with binary and count data. Because binary and count data strongly violate the assumptions of normality and equal variance of ANOVA, the traditional method might lead to unreliable power. In this study, we (1) investigated the influence of binary and count outcomes on statistical power in traditional ANOVA through Monte Carlo simulation, (2) proposed a method for power analysis in ANOVA with binary and count data using generalized linear models, and (3) developed software for power analysis in ANOVA with binary and count data by the proposed method. Practical issues in such power analysis such as the choice of effect size were also discussed.

Symposium 8d: Statistical Power Analysis for Mediation with Non-Normal and Missing Data

Zhiyong Zhang, University of Notre Dame

The existing literature on statistical power analysis for mediation models often assumes that collected data are normal and complete. However, practical data are often non-normal with missing data. This study proposes to estimate statistical power to detect mediation effects based on a two-stage robust method through Monte Carlo simulation. Non-normal data with excessive skewness and kurtosis are allowed in the proposed method. Proportion of missing data can also be specified in calculating statistical power. We illustrate the proposed methods through a simple mediation model, a multiple-mediator model, and a longitudinal mediation model. Software based on the proposed method is also provided and illustrated.

Symposium 8e: Power Analysis for Structural Equation Modeling without Assuming Population Distributions

Ke-Hai Yuan; Zhiyong Zhang, University of Notre Dame

Normal-distribution-based likelihood ratio statistic T_ML is widely used for power analysis in structural equation modeling (SEM). In such analysis, the related discrepancy function F_ML or the corresponding root mean square error of approximation (RMSEA) is often used to specify a non-central chi-square distribution based on which power or sample size is calculated. However, with either violation of normality or not a large enough sample size, both empirical and analytical results indicate that power analysis based on T_ML following the central and/or non-central chi-square distribution is not valid. This talk discusses valid methods for power analysis, including simulation and analytical methods. Different measures of effect size for characterizing the power properties of T_ML are also proposed. Methods for increasing power are proposed and they can improve the power with existing methods substantially.

Item Response Theory- IRT 8: 9:45 AM - 11:15 AM

IRT 8a: Comparing Pattern Scoring with Number-correct Scoring in Mixed-Format Tests

Chen Li, University of Maryland-College Park; Hong Jiao, University of Maryland; Robert W. Lissitz, University of Maryland-College Park

In large-scale standardized educational assessments, scale scores are often used to report examinees' proficiency (Lau, Jiao, & Lam, 2004). Number-correct scoring (NCS) and pattern scoring (PS) are two methods that are widely used in getting scale scores (Yen, 1984). This study investigates the differences between number-correct scoring and pattern scoring in mixed-format tests in terms of examinees' latent ability estimation and performance level classification. Previous study has limited the comparison between PS and NCS in tests with only dichotomous item and used only binary decision in classifying examinees (Lau, Jiao & Lam, 2004; 2006). To address this issue with the prevalence of mixed-format tests in consortium tests that classify examinees into multiple performance levels, such as PARCC, the current study extends the comparison between NCS and PS to mixed-format test and specifically investigates the classification accuracy in categorizing examinees with three performance levels. A simulation study was carried out to compare NCS and PS in mixed-format tests where a 2-PL IRT model was used to calibrate dichotomous items and the generalized partial credit model for polytomous items with 3-score points. Proportion of dichotomous item, sample size, test length and cut scores for classification were manipulated in this study. In total, twenty-four study conditions were simulated with each condition replicated 30 times. Pattern scores were obtained using PARSCALE, and number-correct scores were programmed in R. Study result has showed that PS outperforms NCS in terms of both latent ability estimation and classification accuracy.

IRT 8b: What's the Score? A Psychometric Evaluation of Item-Level Scoring Rules for Educational Tests

Frederik Coomans, KU Leuven and University of Amsterdam

We develop a modeling framework in which IRT models can be developed directly from a scoring rule for dichotomous, polytomous, or even continuous (RT-dependent) items. The models in this framework can be considered normative as they encode how test developers expect test takers to behave in a high-stakes context with an explicitly known scoring rule. We illustrate with an empirical example the fact that these models are often at odds with actual test taking behavior and suggest a number of ways to reduce the discrepancy between actual test taking behavior and the normative IRT models.

IRT 8c: Breaking through the Sum Score Barrier

James Ramsay, McGill University; Marie Wiberg, Umea University, Sweden

When responses to tests or scales are quantified by summing pre-allocated fixed option weights, an important source of information is ignored. This is the variation from item to item in the shape of option and item characteristic curves. That is, sum-scoring uses row information but ignores column information in the response matrix.

We propose a test scoring procedure that has the look and feel of sum scoring but is substantially more powerful. The procedure is illustrated with both real data and simulated data sets. Improvements in root-mean-square error range from 5% to 6% for designed achievement tests to 12% to 14% for typical classroom tests. Moreover, improvements in bias and efficiency for the extremely important cohort at the highest performance level are far higher.

This talk focusses on steps that might lead to the widespread adoption of modern psychometric technology in both large-scale testing and in classrooms.

IRT 8d: Implications of Pairing Strategies and Scoring Approaches in Forced-Choice Assessments

Tim Moses, College Board; YoungKoung Kim, College Board; Carol Barry, College Board

The interest of this study is comparing the scores obtained from different scoring procedures and pairing strategies for tests involving forced choices among pairs of statements. Forced-choice response data will be simulated for several to-be-paired statements measuring one of three intercorrelated traits. The statements are assumed to have available IRT model parameters that came from prior field studies where they were administered in non-paired Likert formats. To address questions about item pairing strategies, the simulations will be conducted such that the statements are paired in different ways with other statements (e.g., based only on difficulty statistics, on difficulty and other judgmental factors, randomly, etc.). To address questions about scoring procedures, the simulated data will be scored based on two procedures: the multidimensional pairwise preference model (Stark, Chernyshenko & Drasgow, 2005) and the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011, 2013). All simulated conditions will use the same set of true values for the three estimated traits. The estimated traits obtained for the three item pairing approaches and the two scoring procedures will be compared to the true values. The implications for designing forced-choice assessments based on item pairing approaches and scoring procedures on estimation accuracy for the traits will be summarized.

IRT 8e: Introduction to an Open Source Java Library for Psychometrics

J. Patrick Meyer, University of Virginia

The psychometrics Java library provides a large set of tools for data analysis and the production of enterprise-grade psychometric software. It is the backbone of jMetrik (Meyer, 2014) and has been ported to general statistical programs such as Stata (Buchanan, 2015). This presentation will provide an introduction to the library's structure, give examples of using its current features, and explain the way it can be extended to new functionality.

The psychometrics library is organized into packages of related features such as reliability estimation, differential item functioning, and exploratory factor analysis. The focus will be on the IRT package and its classes for item response models, marginal maximum likelihood estimation, person scoring, and scale linking. Through a series of short self-contained compilable examples, the libraries features will be explained and demonstrated.

Accessing the library is easy. It is hosted online at GitHub. Anyone can browse the source code, create a branch of it for creating patches or adding functionality, or fork it altogether for creating an independent project. GitHub's version control system retains copies of every version of the library, which allows for easy access to the source code while preventing corruption of the library. More importantly, GitHub is a community of developers. This presentation is aimed at increasing involvement in the psychometrics library community.

Invited Speakers: Steve Reise, Michael Cheung: 11:25 AM - 12:10 PM

Is the Bifactor Model a Better Model or is it Just Better at Modeling Implausible Responses? Application of Iteratively Reweighted Least Squares to the Rosenberg Self-Esteem Scale

Steve Reise, University of Minnesota

Chair: David Thissen

Although the structure of the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) has been exhaustively evaluated, questions regarding dimensionality and direction of wording effects continue to be debated. To shed new light on these issues, we ask: (1) for what percentage of individuals is a unidimensional model adequate, (2) what additional percentage of individuals can be modeled with multidimensional specifications, and (3) what percentage of individuals respond so inconsistently that they cannot be well modeled? To estimate these percentages, we applied iteratively reweighted least squares (IRLS; Yuan & Bentler, 2000) to examine the structure of the RSES in a large, publicly available dataset. Two distance measures for determining case weights were used: (1) reflecting a distance between a response pattern and an estimated model, and (2) reflecting a distance based on individual residuals. We found to be superior to for producing a robust factor pattern and to be more sensitive to, and diagnostic of, improvements in model adequacy as more complex models are fit. A bifactor model provided the best overall model fit, with one general factor and two wording-related group factors. But, based on values, we concluded that approximately 86% of cases were adequately modeled through a unidimensional structure, and only an additional 3% required a bifactor model. Roughly 11% of cases were judged as "unmodelable" due to their significant residuals in all models considered. Finally, analysis of revealed that some, but not all, of the superior fit of the bifactor model is owed to that model's ability to better accommodate implausible and possibly invalid response patterns, and not necessarily because it better accounts for the effects of direction of wording.

Invited Speakers: Mike Cheung: Integrating meta-analysis within structural equation modeling: Theories and applications

Michael Cheung, National University of Singapore

Chair: Cees Glas

Structural equation modeling (SEM) and meta-analysis are two powerful statistical methods in the educational, social, behavioral, and medical sciences. They are often treated as two unrelated topics in the literature. This presentation gives an overview on how popular meta-analytic models, such as univariate, multivariate and three-level meta-analyses, can be integrated under the SEM framework. I will also discuss meta-analytic structural equation modeling (MASEM) that can be used to synthesize findings in SEM.

Item Response Theory- IRT 9: 1:30 PM - 3:00 PM

RT 9a: What if Your Rasch Model Doesn't Fit?

Gunter Maris, CITO-University of Amsterdam

Recent work on the intersection between Item Response Theory models and network models allows for at least three distinct answers to this question. First, one could model the network structure with a freely estimated rank one matrix (known as the 2PL in psychometrics) at a cost of one additional parameter per item. Second, one could model the network as a sparse network in which only a limited number of edges are present (take it equal to the number of items, for fair comparison). Third, one could model the network as being full rank with known eigenvectors (having one eigenvalue to be estimated per dimension). All three of these models have the same number of parameters (2 times the number of items). None of these models will be the true model. Yet, their fit to real data will certainly not be the same. Who wins the competition and why will be the topic of this presentation.

RT 9b: An Assessment of New Item Fit Indices in IRT Models

Jinwang Zou, University of Maryland College Park

Classic item fit indices (e.g., Yen's Q1, McKinley & Mills G2, Orlando & Thissen S) in IRT all require grouping strategy and they have certain limitations in application. For example, when choosing different number of groups, hypothesis tests may provide inconsistent results.

The purpose of this study was to compare the performance of three new item fit indices from Statistics literature by manipulating a variety of factors (e.g., sample size, IRT models, ability distribution and different reasons causing item misfit). The three new fit indices are Standardized Pearson test, Unweighted Sum of Squares test, and Information Matrix test. The main advantage of using these indices is that they are not based on grouping strategy and thus can provide consistent results.

A Monte Carlo simulation was conducted to compare the performance of the three item fit indices. The performance was compared on controlling type I error rate and power. Five possible reasons of item misfit will be explored: DIF, differing item discrimination, omitting guessing parameters, omitting interaction term between two latent traits and omitting quadratic term of latent traits.

Preliminary results showed that the Information Matrix test and Unweighted Sum of Squares test are potentially useful in detecting omitted quadratic term. Both tests have adequate power when sample size is large and quadratic effect is large. However, with medium sample size and medium quadratic effect or small sample size and large quadratic effect, only IM test has adequate power. Other conditions need to be further explored.

RT 9c: Bayesian Fit Analysis for the 1PNO Model

Rudolf Debelak, University of Zurich

There is a wide consensus that the assessment of model fit is an important step in the practical application of models of item response theory. In the literature, numerous numerical and graphical approaches have been described for evaluating the fit of specific models in the last decades. For the Rasch model, previous authors (e.g. Ponocny, 2001) have developed a family of nonparametric model tests, which is based on a Monte Carlo algorithm and suitable for the evaluation of model fit in small datasets. The present study presents and evaluates an application of three tests from this family in the evaluation of the global fit of the one-parameter normal-ogive (1PNO) model in Bayesian item response theory by the application of posterior predictive model assessment. This approach is evaluated by the means of a simulation study, which analyses the Type I error rates and the power of this approach against the violation of parallel item response curves, local independence, and unidimensionality in relation to the size of the analyzed item set and the sample size. Based on the results of the simulation study, it is argued that the presented approach should be regarded as conservative, but has power to detect violations of unidimensionality and local stochastic independence, and, to a lesser degree, violations of equal item discrimination. The evaluated approach may therefore be of interest for the assessment of model fit in practical research.

RT 9d: Limited-Information Goodness-of-Fit testing for Multidimensional IRT Models

Scott Monroe, UMass Amherst

Multidimensional IRT models are becoming increasingly popular for modeling educational and psychological test data. This rise in popularity is due to more tests being designed to measure multidimensional constructs, as well as methodological developments that support estimation, inference, and prediction (e.g., Cai, 2010).

The current research furthers these methodological efforts by proposing a computationally feasible strategy for calculating limited-information goodness-of-fit statistics, such as M_2 (Maydeu-Olivares & Joe, 2006) and C_2 (Monroe & Cai, 2015), for multidimensional IRT models with arbitrary factor structures. These test statistics have proven useful in model evaluation for unidimensional IRT models, as well as multidimensional models with specific factor structures (e.g., bifactor) that facilitate dimension reduction (Cai & Hansen, 2013). However, to the best of our knowledge, the performance of these statistics with arbitrary factor structures (e.g., correlated traits) has not been rigorously studied.

The proposed strategy approximates the test statistic using Monte Carlo methods, thereby avoiding the high-dimensional integration (via quadrature) that would otherwise be necessary. Additionally, a straightforward method is proposed to assess the Monte Carlo error in the estimate.

Preliminary simulation work has been conducted using the proposed strategy to calculate C_2 for a test with 4 correlated dimensions. Results show the Monte-Carlo based C_2 is well-calibrated under the null, and has substantial power to detect an omitted slope.

RT 9e: Bayesian IRT Model Fit Assessment in Measurement Invariance

Xin Xin, University of North Texas

Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) are both common approaches to test Measurement Invariance (MI), and their differences and similarity have long been investigated. Often as prerequisite for group score comparisons, CFA MI procedure evaluates types of equivalence from configural, metric, to scalar (Vandenberg & Lance, 2000), and each type is tied to substantive meaning (Millsap & Hartog, 1988). While mathematically equivalent to binary CFA, two-parameter (2PL) IRT has computational advantages (Kamata & Bauer, 2008). Applying CFA MI procedure to 2PL IRT models may take advantage of both the substantive interpretation and computational power. In recent years, Bayesian IRT model fit assessment keeps attracting more attention. Posterior Predictive Checking (PPC; Rubin, 1984) is a popular Bayesian model checking tool because of its intuitive interpretation and strong theoretical basis (Sinharay, 2006). The present study implemented CFA steps on unidimensional 2PL IRT models, and adopted PPC method to diagnose both item-fit and model-fit using discrepancy measures proposed by Hoijtink (2001) and Sinharay, Johnson, and Stern (2006). Meanwhile, examples of types of CFA MI equivalence were generated in order to compute Root Mean Square Error (RMSE) and bias of item estimates. Prevailing item fit and model fit statistics were also provided, such as information criteria, the Pearson X2 statistic, and the likelihood ratio statistic G2 (Orlando & Thissen, 2000), as conveniently offered by flexMIRT (Cai, 2012).

Symposium 9: Current Advances in Confidence Interval and Prediction Interval Estimation: 1:30 PM - 3:00 PM

Symposium 9a: Constructing Confidence Intervals about the Mean of Non-Normal Distributions

Jolynn Pek, York University; Octavia Wong, York University; Augustine C. M. Wong, York University

There is a growing emphasis on constructing and reporting confidence intervals (CIs) in psychological research, which is fueled by the best practice movement (e.g., Cumming, 2014; Funder, et al., 2009; Wilkinson and the Task Force for Statistical Inference, 1999). When researchers encounter a non-normal distribution, there are several approaches to construct a CI about the mean. The two most popular approaches are to (a) transform the sample distribution to obtain a normal distribution and construct a CI, and (b) ignore non-normality of the sample distribution and construct a CI. Different types of CIs (Wald-type, bootstrap) for either approach can be constructed. First, we highlight issues involved with transformations related to interpretability and parameter invariance. Second, we present a simulation study which empirically accesses the properties of these CIs (coverage, efficiency). Finally, we make recommendations on how best to construct CIs in practice.

Symposium 9b: Large-sample Sampling Variability for Two Differential Test Functioning Measures

R. Philip Chalmers, York University; David B. Flora, York Unviersity; Alyssa Counsell, York University

Differential test functioning (DTF) occurs when one or more items in a test contain differential item functioning (DIF) and the aggregate of these effects are witnessed in the expected test score functions. In many applications, DTF can be more important than DIF when the overall effects of DIF at the test level can be quantified. However, statistical methodology for detecting and understanding DTF and its associated variability has generally been underdeveloped. This talk presents two distinct semi-parametric DTF measures which are useful for quantifying marginal and conditional bias effects, and demonstrates that their respective sampling variability can be expressed through the sampling variability of the item parameter estimates. Large-sample confidence intervals and hypothesis tests for the presented DTF measures are introduced, and results based on Monte Carlo simulations are presented.

Symposium 9c: Bootstrap-Calibrated Prediction Intervals for Response-Pattern Scores in Item Response Theory

Yang Liu, University of California, Merced; Ji Seung Yang, University of Maryland, College Park

When the parameters in the scoring IRT model need to be estimated, inferences about the score of a given response pattern are often made from the corresponding posterior distribution of the latent traits evaluated at the maximum likelihood estimates (MLE) of model parameters, i.e., the plug-in posterior. Despite its popularity, the plug-in method ignores the sampling variability of the MLE and overstates the precision of the resulting scores, especially when the calibration sample is not large. The issue of incorporating the sampling error of the MLE into the inference about scale scores has drawn a substantial amount of attention (e.g., Mislevy et al., 1993; Cheng and Yuan, 2010) over years. In the current work, scoring is deemed as making predictive inferences about a plausible value generated from the true posterior. We propose to construct interval estimates of response-pattern scores based on a bootstrap-calibrated predictive distribution (e.g., Beran, 1990; Fonseca et al., 2014), which are expected to have more accurate coverage compared to the interval estimates derived from the plug-in posterior. A simulation study is conducted to compare the bootstrap calibration method with various existing candidates that have been demonstrated to outperform the plug-in method in small samples, such as the multiple-imputation approach (Yang et al., 2012) and the full Bayesian approach (e.g., Patz & Junker, 1999).

Symposium 9d: Bootstrap Confidence Intervals of Fit Indices by Inverting a Test

Chuchu Cheng, Boston College

We propose a new method of constructing confidence intervals (CIs) for fit indices in SEM. Currently, the newest method in bootstrapping CIs for fit indices is based on the quantiles of a bootstrapping sampling distribution at a single fixed level of misspecification. This is parallel to a Wald type CI, where the quantiles of a single sampling distribution produced from the estimated parameter value are used, but in contrast to a profile-likelihood-based CI, whose boundary produces a designated p-value in a likelihood ratio test. Because the profile-likelihood approach is usually superior to a Wald type CI, we propose to construct CIs by searching for levels of misspecification that produce an appropriate p-value and expect our method to perform better than the existing approaches.

Symposium 9e: A Simultaneous Confidence Interval of Effect Sizes of Pairwise Comparisons

Hao Wu, Boston College

When psychology has been moving from testing an exact null hypothesis to using confidence intervals of an effect size, simultaneous CIs for effect sizes are still under-developed. I will present such an effort for pairwise comparisons. The set of simultaneous CIs for k groups is calculated by searching for the appropriate size of a region of given shape in the (k-1)-dimensional space of contrasts so that the p-value from a multivariate non-central t distribution equals to a given alpha level.

Longitudinal Data Analysis- LDA 1: 1:30 PM - 3:00 PM

LDA 1a: Comparison of Estimators for Single-Case and Multilevel AR(1) Models

Casper J. Albers, University of Groningen; Tanja Krone, University of Groningen; Mariejke E. Timmerman, University of Groningen, Heymans Institute for Psychological Research, Psychometrics and Statistics

There is a steep rise of longitudinal studies within psychology that employ the AR(1)-model. This holds for both single case as multi-subject designs. Data such as Ecological Momentary Assessment studies usually have (much) fewer time points than the economic data sets that lie at the foundations of the AR(1)-model. It is therefore important to study to what extent the methods are as appropriate for psychological data as they are for economic time series.

We examined the comparative performance of the available AR(1) estimators for short time series through simulation studies. The first study focuses on the single case design. A comparison is performed between four frequentist approaches and two Bayesian approaches. From the frequentist approaches, the maximum likelihood estimator performed best, and from the Bayesian approaches, the one with the symmetrised-reference prior. The second study investigated the robustness of estimators against model misfit, by simulating data from an ARMA(1,1) model and estimating parameters with an AR(1) model.

The third study focuses on multi-subject designs. For the multilevel AR(1) model, we performed a comparison between the two estimators that performed best for the single subject design. Furthermore, we examined the difference between the 'fixed' and 'random' modelling approach. We found that the random estimators show less bias and higher power, compared to the fixed estimators.

Finally, implications of the results for (a priori) power analyses of AR(1) models will be discussed.

LDA 1b: Estimating Relationships in Change for Latent Variables Using Difference Factors

Holmes Finch, Ball State University; Serena Shim, Ball State University

Social scientists frequently want to assess how change over time in one construct (e.g. achievement, motivation, personality) is related to change in another construct. Such questions are often addressed using growth curve models (GCMs). However, when observed indicators are measured at only two time points, there are not sufficient degrees of freedom to fit GCMs. Longitudinal panel models can also be used to investigate relationships between constructs at multiple times points, but do not readily assess how change in one construct relates to change in another.

This study investigates a new latent difference factor model for estimating relationships in change over time for two or more latent variables. For example, consider two factors measured by multiple observed indicators taken at two time points. The researcher wants to know whether change from time 1 to time 2 in one factor relates to change in the other over the two times. To estimate latent difference factors, difference scores between the two times are calculated for each indicator. These difference scores then serve as indicators for difference factors for each latent construct. The relationship between the two is then examined using a structural equation model. If the difference factors are found to be related, the researcher concludes that change in one factor correlates to change in the other.

A simulation study demonstrates that the factor difference model accurately estimates relationships in change over time for two factors. Simulation results will be reported in detail, and the model will be demonstrated using real data.

LDA 1c: Handling Sparse and Missing Data Using Functional Mixed Effects Models

Kimberly Ward, Arizona State Universtiy; Hye Won Suk, Arizona State University

Researchers who collect longitudinal data often encounter the issues of sparse and missing data. These issues can hinder the estimation of individual growth trajectories, a major component of longitudinal research. We propose treating longitudinal data as functional data, and analyzing it by means of a functional mixed effects model (FMEM). Functional data are defined as curves evaluated at a finite number of time points, which are assumed to reflect underlying smooth processes of interest. The utility of this method lies in estimating the underlying smooth curves from discrete, raw data using non-parametric smoothing methods. The FMEM framework uses information from all individuals to estimate both mean and individual curves. A series of two simulation studies are presented investigating the ability of FMEMs to estimate the mean and individual trajectories under various conditions such as sparseness, type of missingness, shape of trajectory, percent missingness, and irregularity of time points. Results reveal FMEMs are robust to sparseness and type of missingness. More specifically, results showed accurate estimation with as few as five time points, missing data rates as high as 50%, and systematically missing data. Estimation accuracy improved as the number of time points per curve increased and as the percent of missing data decreased. The amount of improvement in estimation accuracy as the number of time points increased was dependent on the trajectory shape. The use of random time points was shown to improve estimation accuracy further. These results offer researchers a method of analysis robust to sparseness and missing data.

LDA 1d: Bayesian VAR(1)-Modeling to Unravel Emotion Dynamics

Mariejke E. Timmerman, University of Groningen, Heymans Institute for Psychological Research, Psychometrics and Statistics; Tanja Krone, University of Groningen; Peter Kuppens, Leuven University; Casper J. Albers, University of Groningen

Emotion dynamic research typically aims at revealing distinct information on affective functioning and regulation. Herewith, one distinguishes various elementary emotion dynamic features, which are studied using intensive longitudinal data. Typically, each emotion dynamic feature is quantified separately, which hampers the study of relationships between various features. This complicates the validation of theories stating relations between these features. In emotion research, the length of the observed time series is limited, and often suffers from a high percentage of missing values. In this paper we propose a Bayesian vector autoregressive model (VAR), a variant of the Bayesian dynamic model, that is useful for emotion dynamic research. The model encompasses the six central emotion dynamic features at once, and can be applied with relatively short time series, including missing data. The individual emotion dynamic features covered are: long and short term variability, granularity, inertia, cross-lag correlation and the intensity. The model can be applied to univariate and multivariate time series, allowing to model the relationships between emotions. Further, it may model multiple individuals jointly. One may include external variables. The model can be specified for non-Gaussian observed data, and can deal with missing data through a link function. We illustrate the usefulness of the model with an empirical example using relatively short time series (47 to 70 measurements) of three emotions, with missing time points within the series, measured for three individuals. We discuss how the model could be extended, and the limitations one could face when estimating the model for empirical data.

LDA 1e: Estimation of Time-Unstructured Nonlinear Mixed Effects Mixture Models

Sarfaraz Serang, University of Southern California; Kevin J. Grimm, Arizona State University; John J. McArdle, University of Southern California

Change over time often takes on a nonlinear form which can introduce complexities in model estimation. Furthermore, these change patterns can sometimes be characterized by heterogeneity due to underlying unobserved groups in the population. Nonlinear mixed effects mixture models provide one way of addressing both of these issues simultaneously. The purpose of this study is to extend this class of models to accommodate individually varying measurement occasions. We develop methods to fit these models in both the structural equation modeling framework as well as the Bayesian framework and evaluate the performance of these methods in an attempt to provide researchers with some practical recommendations regarding their use. Simulation results show that the main force driving the success of these methods is the separation between latent classes. When these classes are well separated, even a sample of 200 individuals appears to be sufficient. Otherwise, a sample of 1000 or more may be required before parameters can be accurately recovered. Ignoring heterogeneity in time of measurement also led to substantial bias, particularly in the random effects parameters. Finally, we demonstrate the application of these techniques to an empirical dataset involving the development of reading ability in children.

Equating- EQUATE 2: 1:30 PM - 3:00 PM

EQUATE 2a: Comparing Equating Performance for Various Synthetic Populations

Hyung Jin Kim, University of Iowa; Won-Chan Lee, University of Iowa

For large-scale testing programs, multiple forms of a test are constructed as parallel as possible, and, sometimes, are administered at different test dates due to test security or other practical concerns. In such cases, the common-item nonequivalent groups (CINEG) design often is used with a set of items in common between new and old forms. For the CINEG design, some equating methods use a single synthetic population based on new and old populations.

Choices over proper synthetic populations have been an issue. Brennan and Kolen (1987) suggested assigning a full weight to a new group for conceptual simplifications, whereas Angoff (1987) suggested equal weights, concerning that different weights might yield population parameters that are biased. However, there has not been much research that examines whether equating performance depends on how a synthetic population is defined.

This study examines how equating performs differently depending on the composition of a synthetic population. Proportions of common items, standardized differences in common-item scores between new and old groups, equating order (i.e., whether or not a new group has a higher ability level than an old group), sample sizes, and form differences in difficulty are considered as study factors. A preliminary study shows that, when the new form is easier than the old form, choices over proper synthetic populations are different from those for the opposite case. Moreover, it shows that choices over proper synthetic populations depend on the interaction between equating methods and equating order.

EQUATE 2b: The Epanechnikov and Adaptive Continuization Methods in Kernel Equating
Jorge González, Pontificia Universidad Católica de Chile; Alina A. von Davier, Educational Testing Service, USA

Gaussian kernel continuization of the score distributions has been the standard selection in kernel equating. Recently, Lee & von Davier (2011) have explored the use of other kernels such as the logistic and uniform. When kernel smoothing is used to estimate probability distribution functions, it is known that “boundary bias” problems arise. In the context of kernel equating, Cid & von Davier (2014) explored the use of two alternative kernels (the Epanechnikov and Adaptive kernels) for the estimation of score densities, as potential alternative approaches to reducing boundary bias. In this talk we illustrate the use of both the Epanechnikov and Adaptive kernels in the actual equating step using the R-package SNSequate (González, 2014). A comprehensive comparison is presented here: the two new kernel equating methods are compared with each other and with the Gaussian, Logistic and Uniform kernels both with respect to their potential for continuizing the test score densities and to the equating results.

EQUATE 2c: Accumulative Linking Error in Trend Measurement in Large-Scale Assessments
Lauren Harrell, National Center for Education Statistics

Accumulative linking error may impact the precision of trend comparisons across years when chained linkages are used. Trend comparisons, in which the means and percentiles of group and subgroup distributions are compared across time, are major targets of inference in the reporting of large-scale group-score assessments, such as the National Assessment of Educational Progress (NAEP). Under current operations, each assessment is linked with only the immediately preceding assessment year within grade and subject area through common-item nonequivalent-groups design via concurrent item response theory calibration. When a hypothesis test is conducted to compare a the mean scale score of a given assessment year to the first year of that assessment, the distributions of scale scores have gone through a number of chained linkages to arrive at that hypothesis test. The potential for accumulative linking error across chained linkages and subsequent impacts on trend comparisons are demonstrated using simulations under designs similar to NAEP, and methods for adjusting for the decrease in precision over time are presented.

EQUATE 2d: Ensuring Quality Over Time by Monitoring the Equating Transformations
Marie Wiberg, Umea University, Sweden

When administering high-stake tests consecutively over several administrations it is of great importance to ensure quality over time at all stages. One important part of ensuring the quality is to make sure that the equating procedure work as intended, especially when the composition of the test taker groups might change over the administrations. The aim of this paper was to examining the equating transformations over a large number of administrations of a college admission test which is given twice a year and where the test takers can use the results to apply for university during a period of five years. The test has an external anchor and thus a non-equivalent anchor test design is typically used to equate the test. Different equating methods, different test groups and different linking plans are examined and discussed. The used methods are traditional equating methods, kernel equating methods, and item response theory methods. The preliminary results suggest that the composition of the test taker group within an administration has impact, and different linking strategies and different equating methods give somewhat different results. The importance of assuring that similar conditions are met over time and how one can take care of different composition of test takers within the different administrations are discussed. Recommendations for performing equating consecutively over several administrations in the future are given.

Multilevel/Hierarchical/Mixed Methods- MLM 3 : 1:30 PM - 3:00 PM

MLM 3a: Item Response Modeling of Count Data with Inflation and Heaping

Brooke E. Magnus, Marquette University; David Thissen, University of North Carolina at Chapel Hill

Questionnaires that comprise multiple items eliciting count responses are becoming increasingly common in psychology and health research. Often, these surveys are designed to assess the severity of symptoms and ask respondents to recall the frequency of various thoughts or behaviors over a pre-specified period of time. Retrospectively-reported item response data from these types of surveys pose several analytic challenges, including inflation at zero and the maximum that may represent distinct subpopulations, as well as heaping at preferred digits (i.e., counts that are multiples of five); such data complexities are not well-suited for conventional IRT modeling approaches and software. Simply modifying a traditional IRT model to invoke a log link function and a Poisson or negative binomial conditional response distribution is not likely to account for the potential subpopulations and individual differences in response style that are observed in retrospectively-reported count data. This research addresses this problem by combining methodological approaches from three related but distinct literatures: IRT models for multivariate count data (e.g., Wang, 2010), latent variable models for heaping and extreme responding (e.g., Wang & Heitjan, 2008; Bolt & Johnson, 2009), and mixture IRT models (Finkelman et al., 2011; Wall et al., 2015). First, we present the theoretical framework for the latent class IRT model. Then, we carry out a small simulation to test the implementation of parameter estimation in R. Finally, we apply the model to empirical data from the Behavioral Risk Factor Surveillance System.

MLM 3b: Detecting Multiple Random Knots in Bayesian Piecewise Growth Mixture Models

Eric F. Lock, University of Minnesota; Nidhi Kohli, University of Minnesota

Piecewise growth mixture models (PGMM) are a flexible and useful class of methods for analyzing segmented trends in individual growth trajectory over time, where the data come from a mixture of two or more latent classes. These models allow each segment of the overall developmental process within each class to have a different functional form; examples include two linear phases of growth, or a quadratic phase followed by a linear phase. The knot is the time of transition from one developmental phase (segment) to another. Inferring the location of the knot(s) is often of practical interest, along with inference for other model parameters. A random knot allows for individual differences in the transition time within each class. The primary objectives of our study are: (1) to develop a PGMM using a Bayesian inference approach that allows the estimation of multiple random knots within each latent class; (2) to develop a procedure to empirically detect the number of random knots within each latent class; and (3) to empirically investigate the accuracy and precision of the model parameters, including the random knots, via Monte Carlo simulation study. Finally, we will use two real data examples to illustrate the utility of this method.

MLM 3c: A Three-Level Multilevel Growth Model with Multiple Measures at Level 1

Yasuo Miyazaki, Virginia Tech University

A three-level multilevel growth model is considered by conceptualizing that items are nested within measurement occasions, and in turn, measurement occasions were nested within people. Conceptually, the, lower two level (level 1 and level 2) model represents a measurement model and the upper two level (level 2 and level 3) model represents a growth model, and those measurement and growth models are combined into a single three-level model. Though the proposed model increases a complexity of the model, there are several advantages of using this model over a standard two-level growth model. First, the proposed model can be used to measure the change better than the traditional one since measurement error will be taken into account. Second, this model can be fitted to the data that have only two waves of information as far as there are multiple measurements available for each wave. It is typically considered that the acceptable growth model can be fitted with the data that have minimum of three waves, but there are many occasions in which only two waves are available. The proposed model provides a model that accommodates the necessity of measuring the change more accurately with only two waves information. An illustrative example using an anxiety data will be provided to facilitate understanding the advantages.

MLM 3d: Investigation of Level-1 Error Covariance Structure in Multilevel Growth Analysis

Yi Lu, American Institute for Research

Multilevel longitudinal analysis offers great flexibility in modeling the covariance structure for both between-subject fixed effects, random effects, and within-subject random errors. In multilevel longitudinal analysis, the assumption of i.i.d.~(0,) of within-subject residuals (level-1 covariance matrix,) may bias estimates of the parameters of interest since it is common that residuals are autocorrelated. There are some studies investigated level-1 covariance structures in multilevel growth curve models focusing condition of homogeneous variance and/or equal measurement interval. (e.g., Hoffman, 2015; Murrphy &Pituch, 2009). Fewer studies investigated optimal level-1 covariance structure in the condition of heterogeneous variance and/or unequal measurement points. This study investigates the following research questions:

1. What's the optimal under the assumptions of homogeneous/heterogeneous error variance?
2. What's the optimal under the condition of equal/unequal time intervals?
3. What's the optimal under the cross condition of homogeneous/heterogeneous error variance and equal/unequal time intervals?

This study discusses multilevel growth modeling in the framework of hierarchical linear modeling (HLM). The analysis is conducted using SAS PROC MIXED (SAS Institute Inc., 2008). Different candidate covariance matrices are proposed and tested by patterns of variances, time intervals, and cross conditions of variances and time intervals. When selecting the covariance structure, this study considers a variety of standards such as number of parameters, the interpretation of the covariance structure, diagnostic results, deviance statistics and information criteria, etc. The effects of different error covariance specification on fixed and random effects, intra-class correlation (ICC), etc. will be discussed.

Classical Test Theory- CTT 1: 1:30 PM - 3:00 PM

CTT 1a: Dimensionality: Cronbach's View Versus a Factor Analytic View

Ernest C. Davenport, Jr., University of Minnesota; Mark L. Davison, University of Minnesota

In Cronbach's seminal article in which he introduced coefficient alpha as a measure of reliability (Cronbach, 1951), he also operationalized dimensionality as the proportion of variance attributable to the first principal factor running through the test. Cronbach's operationalization, however, differs from a factor analytical view of dimensionality. The latter view focuses on the number of primary factors necessary to reproduce the variance/covariance (or correlation) structure of the items in question. The proposed paper contrasts these two definitions of dimensionality as it examines what it means for a test to be multidimensional. The paper will begin with a demonstration that shows for a factor analytic multidimensional test, Cronbach's definition would suggest an effectively unidimensional test as item complexities and/or factor correlations increase. Note that item complexities and factor correlations suggest that the items have something in common at a higher order. This suggests that factor analytic dimensionality is more a function of the primary factors while Cronbach's definition is related to the extent that the items have a consistent higher order factor. This proposed research has implications for interpretation of multidimensional tests. Given that in his 1951 article Cronbach says, “The proportion of the test variance due to the first factor among the items is the essential determiner of the interpretability of the scores.” suggests that scores for factorially multidimensional tests may be interpretable to the extent that the items lead to a higher order factor.

CTT 1b: Overestimation of Reliability by Guttman’s Coefficients and the GLB
Klaas Sijtsma, Tilburg University; Pieter Oosterwijk, Tilburg University, The Netherlands; Andries van der Ark, University of Amsterdam

For methods using statistical optimization to estimate test-score reliability, we investigated bias and sampling error, and possible overestimation of test-score reliability. Optimization methods do not only exploit real relationships between items but also tend to capitalize on sampling error and do this more strongly as sample size is smaller. The optimization methods were Guttman’s coefficients lambda4, lambda5, and lambda6, and the greatest lower bound to the reliability (GLB). Coefficient lambda2 has been suggested in the literature as the preferred reliability lower bound and was included in the study as a benchmark. Bias and sampling error, and possible overestimation of test-score reliability were investigated in a simulation study for varying sample size, test length and data dimensionality. We found that coefficient lambda4 and the GLB often overestimated test-score reliability. When sample size exceeded 250 observations, coefficients lambda2, lambda5, and lambda6 provided reasonable to good statistical results, especially when data were two-dimensional. In most conditions, coefficient lambda2 produced the best results.

CTT 1c: Interval Estimation Procedures for Polytomously-Scored Items
Kyung Yong Kim, University of Iowa; Won-Chan Lee, University of Iowa

Assuming that a sample of n polytomously-scored items is drawn at random from an undifferentiated universe of such items, measurement errors can be modeled using a multinomial distribution. Under the multinomial error model, an examinee's observed scores obtained from repeated measurements conditional on true category-proportion correct scores follow a multinomial distribution. The purpose of the present study is twofold: (1) propose normal approximation and Wilson's score-like interval estimation procedures using a multinomial error model and (2) compare these procedures to the general method for constructing ad hoc "exact" (Lee, 2005) interval estimates through a simulation study under various study conditions. Across all the simulation conditions, the Wilson's score-like interval estimation procedure return intervals with actual coverage probabilities much closer to the nominal level than the other two procedures. In spite of the better coverage probabilities, the interval widths are not necessarily wider. Another attractive aspect of the Wilson's score-like procedure is that it has a closed-form expression that can be easily evaluated.

CTT 1d: Decision Consistency and Classical Reliability
Won-Chan Lee, University of Iowa; Stella Kim, University of Iowa; Robert Brennan, Unversity of Iowa

Decision consistency refers to the extent to which examinees are classified into the same performance categories based on two parallel instances of a measurement procedure. A conventional wisdom might suggest that a more reliable test lead to higher decision consistency. This study explores the relationship between classical reliability and decision consistency, attempting to answer whether high reliability always leads to high decision consistency, and whether the relationship remains the same regardless of cut-score locations. We also propose a framework that can be used to predict decision consistency for lengthened or shortened test forms with the assumption that the cut score changes proportionally according to the test length. Two psychometric models are employed to address the aforementioned study questions: the beta-binomial model (Huynh, 1976) and the normal approximation method (Peng & Subkoviak, 1980). The results based on both models suggest that, for a fixed test length, the relationship between classical reliability and decision consistency does not necessarily show a monotone increasing pattern—the pattern varies as the cut-score location changes. The predicted decision consistency increases as the test length increases, and the amount of increase depends on the location of the cut score. Some practical usage of the proposed prediction model is discussed.

Presidential Address: 3:15 PM - 4:15 PM

Presidential Address: Differential Item Functioning from a Multidimensional Item Response Theory Perspective

Terry Ackerman, The University of North Carolina at Greensboro

Chair: Anders Skrondal

Differential item functioning, DIF, can best be understood from a multidimensional item response theory perspective. This talk will be begin with a brief background detailing the compensatory two-parameter MIRT model and how items are represented. This will be followed with an explanation of the reference composite, the resulting unidimensional scale that is created when practitioners try to fit a unidimensional IRT model to a two-dimensional latent space. Using this as a foundation, how DIF can occur will be discussed, including the affects of conditioning scores and differences in the underlying latent ability distributions. The talk will end with a quiz in which the audience will be called upon to identify which group is being favored using DIF results from an actual high stakes standardized test.