<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00-9:30am</td>
<td>Continental Breakfast</td>
</tr>
<tr>
<td>9:30-9:40am</td>
<td>Welcome and Introductory Remarks: Jeffrey R. Harring</td>
</tr>
<tr>
<td>9:40-10:40am</td>
<td>Session 1: Longitudinal Methodology</td>
</tr>
<tr>
<td></td>
<td><strong>Chair: Kaiwen Man</strong></td>
</tr>
</tbody>
</table>
|                  | Oh %&$#! They Cut My Funding: Using Planned Missing Data Methods to Salvage Longitudinal Research  
|                  | Yi Feng and Gregory R. Hancock                                        |
|                  | A Look at General Health with a Second-Order Latent Growth Model      |
|                  | Frank A. Rojas                                                       |
|                  | Longitudinal Randomized Controlled Trials with Item Response Data    |
|                  | Marian M. Strazzieri and Ji Seung Yang                               |
|                  | Design and Analytic Implications in Modeling Student Mobility Across Correlated Schools  
|                  | Tessa Johnson, Yi Feng, and Laura M. Stapleton (with special thanks to Dr. Sweet)       |
| 10:40-11:00am    | Break                                                                |
| 11:00-12:00pm    | Session 2: Methodological Developments in Measurement                |
|                  | **Chair: Evan Olson**                                                |
|                  | Mining Process Data to Detect Item Harvesters                        |
|                  | Manqian Liao, Jeffrey Patton, Ray Yan (Financial Industry Regulatory Authority), and Hong Jiao |
|                  | An Evaluation of Hierarchical Models Relating Item Response Format, Accuracy, and Speed   
|                  | Xin Qiao                                                             |
|                  | Measurement Invariance of the DINA Model: Implications for Detecting Problem-Solving  
|                  | Jung-Jung Lee                                                        |
|                  | The Performance of $l_\gamma$ under IRT Model Misspecification and Cheating  
|                  | Chen Tian                                                            |
12:05-1:15pm  Lunch—in Benjamin: EDU 1315 (the Art room on the first floor)

1:15-2:00pm  Session 3: Statistical Methodology and Applications  
              Chair: Tiago Caliço

              A Brief Introduction to Markov Chain Monte Carlo  
              Christian Meyer

              Standard Error Adjustment for Projected IRT Scores in a Fixed-Item-Parameter  
              Linking Design  
              Reni Xu

              The Influence of Poor Measurement Quality on Model Selection: An Example Using  
              the GRIT Scale  
              Weimeng Wang

2:05pm  Concluding Comments: Gregory R. Hancock
Oh %&$#! They Cut My Funding: Using Planned Missing Data Methods to Salvage Longitudinal Research  
Yi Feng and Gregory R. Hancock

In longitudinal studies, any unexpected funding cut can create huge challenges for applied researchers. When it occurs, the principal investigator usually needs to articulate a new plan showing that the study’s goals can still be met (i.e., with enough power to detect the effect of focal interest) with the reduced budget, or otherwise the remainder of the funding will be rescinded altogether. Planned missing (PM) designs can be a very promising solution in such a scenario, which allow researchers to complete the study at a reduced cost but without necessarily compromising the power for testing focal parameters. Although PM designs have shown great potential in previous simulation studies, the implementation remains quite limited, as it requires copious amounts of work and careful planning for applied researchers to come up with an optimal, yet practical, design. The goal of the current study is to introduce an R package, simPM, which automates a simulation-based search for PM designs in the context of longitudinal studies with a direct application to the scenarios where an unexpected funding cut occurs during the course of a study. Illustrative examples are used to demonstrate how this tool can greatly assist applied researchers by automatically comparing across different designs and searching for an optimal design with sufficient power for testing the focal parameters.

A Look at General Health with a Second-Order Latent Growth Model  
Frank A. Rojas

The health sciences has traditionally modeled BMI longitudinally with mixed-effects models. Depending on the research question and design, it can depict change or growth adequately. However, according to Box (1976), "all models are wrong" therefore other models exists that can depict change or growth. One such model is the second-order latent growth model. A nationally representative longitudinal data set, the National Longitudinal Survey of Youth 1997 (NLSY97) obtained from the Bureau of Labor Statistics, was used to illustrate the model building process and how the model can answer a different type of research question that would not have been possible using a mixed-effects model. Issues with model fit will be discussed and how these issues can be remedied. The importance of measurement invariance will be highlighted, as this is the major benefit of using a second-order latent growth model over mixed-effect models.

Longitudinal Randomized Controlled Trials with Item Response Data  
Marian M. Strazzeri and Ji Seung Yang

Longitudinal randomized controlled trials (RCTs) are often used to estimate treatment effects on outcomes that are typically measured by administering the same items repeatedly over time. To analyze item response data collected through such a design, a multiple-group latent growth model (MG-LGM) is theoretically the best choice because measurement and structural parameters are estimated simultaneously. However, full information maximum likelihood (FIML) estimation of a MG-LGM with
categorical indicators is computationally challenging due to the “curse of dimensionality.” As a result, many applied researchers adopt a “divide and conquer strategy” in which the measurement and structural models are fit separately (i.e., multi-stage estimation). Motivated by an empirical study (patient-reported physical functioning in a cancer clinical trial) and methodological literature (Glas, Geerlings, van de Laar, & Taal, 2009), the current study aims to review available single- and multi-stage approaches to modeling latent growth measured by Likert-type response scales in RCTs. We assess the performance of five computationally more tractable approaches across varying conditions (sample size, effect size, level of informative sample attrition) in terms of bias in estimated parameters and corresponding standard errors, power, and Type I error rate via a Monte Carlo simulation study. The five methods reviewed include multi-stage estimation using (1) expected a posteriori (EAP) scores, (2) plausible values (PVs; Mislevy, Beaton, Kaplan, & Sheehan, 1992), (3) and doubly-imputed plausible values (DIPVs; Yang, Hansen, & Cai, 2012), (4) one-stage FIML estimation of a mis-specified form of the LGM that omits common item effects (CIEs; Zheng & Yang, 2018); and (5) one-stage limited information estimation of the fully-specified LGM. Results suggest multi-stage estimation with unidimensional calibration is a viable alternative when we have measurement invariance both over time and between groups, and that one-stage estimation should only be utilized when the model is correctly specified and data are not missing informatively.

**Design and Analytic Implications in Modeling Student Mobility Across Correlated Schools**

*Tessa L. Johnson, Yi Feng, Laura M. Stapleton (with special thanks to Tracy M. Sweet)*

Longitudinal, multilevel education studies provide a wealth of information with implications for program evaluation and policy. However, these data are often quite complex, and little empirical evidence exists to guide the design and analysis of such studies. Current best-practices for analysis of these data structures make strong assumptions about the purity of hierarchical clustering and the lack of correlation among schools attended by mobile students. In the present study, we violate these assumptions and investigate the effect of student mobility across correlated schools on the parameter estimates of two modeling approaches: (1) a multiple membership random effects model (MMREM), and (2) a first-school hierarchical linear model (HLM). In addition, we examine data from a statewide longitudinal database to provide researchers with a gauge for the magnitude of the correlation among schools attended by mobile students, which may aid in the design of future studies and analysis plans.

**Mining Process Data to Detect Item Harvesters**

*Manqian Liao, Jeffrey Patton, Ray Yan (Financial Industry Regulatory Authority), and Hong Jiao*

Item harvesters who memorize, record and share test items can jeopardize the validity and fairness of credentialing tests as well as cause financial and reputational losses to testing companies. The difficulty in delineating a behavioral pattern of item harvesting makes it challenging to detect item harvesters. This study develops an approach that is based on k-means clustering, an unsupervised learning algorithm, to discover the behavioral patterns of the majority of examinees (i.e., behavioral archetypes) and identify examinees whose behaviors deviate from the archetypes, which serves as a starting point to detect item harvesters. The proposed approach can incorporate a variety of process data, making it possible to capture more subtle differences between item harvesters and normal examinees. Data from a high-stakes and high-volume licensure exam are used to demonstrate the proposed approach. Implications and future directions are discussed.
An Evaluation of Hierarchical Models Relating Item Response Format, Accuracy, and Speed

Xin Qiao

The current study evaluates innovative item types for computer-based tests using models for speed and accuracy in a confirmatory factor analysis/structural equation modeling framework using empirical assessment data. Specifically, the interest is on the impact of different response formats for selected-response (SR) items on model parameters for accuracy and speed. Several studies have evaluated the innovative item types in terms of their psychometric properties. However, in none of these studies, response time was included in the modeling approach. In addition to investigate the impact of response format on item difficulty, discrimination, reliability, and information per unit time, we also examined the impact of response format on dimensionality with respect to accuracy and speed in the hierarchical modeling framework. Furthermore, we evaluated the differences in dependencies within and between items with respect to response format, accuracy, and speed. The results indicate that item features such as response format are related to item parameters for both accuracy and speed.

Measurement Invariance of the DINA Model: Implications for Detecting Problem-Solving

Jung-Jung Lee

The utility of Cognitive Diagnostic Models (CDMs) have received more and more attention from researchers because of the models’ unique characteristics in providing diagnostic feedback to individual students on their learning. In large-scale international assessment programs such as PISA, students from different countries participate in the assessment. Measurement invariance across countries may often be questioned. Differential item functioning (DIF) is often utilized to detect item level invariance. Assuming students from different countries use different strategies to solve problems, Q-matrix developed based on one country might be misspecified for students from the other country, especially for items that require more attributes. The current study aims to investigate the parameter invariance property of the DINA model when the Q-matrix is misspecified and DIF is present due to different problem-solving strategies, using simulation. Multiple variables are manipulated to simulate different study conditions including sample size, DIF magnitude, and DIF type. Three DIF detection method including the Mantel-Haenszel procedure, the Simultaneous Item Bias Test (SIBTEST), and the Wald Test are explored in detecting DIF. The performance of each DIF detection method is evaluated in terms of Type I error rate and power.

The Performance of \( I^* \) under IRT Model Misspecification and Cheating

Chen Tian

The person fit statistic \( I^* \) reflects the deviation between the observed item-score patterns and model predictions; however, it provides little information about the type of misfit. This study examines the performance of \( I^* \) when person misfit comes from the IRT model misspecification. Model misspecification happens when 1PL/2PL models are fitted while the data generating model is 3PL. Responses to a 40-item linear test were simulated, in which 20% responses coming from 3PL model was fitted by 1PL or 2PL model. Results show that the detection rates of \( I^* \) were close to the \( \alpha \) level, especially when the less-misspecified 2PL model was used. This study also examines the detection rate of \( I^* \) for item pre-knowledge and answer copying, with a varying percentage of tampered items and the degree of cheating. Results show that a response pattern with a higher percentage of tampered answers or degree of cheating is more likely to be detected, especially for low ability examinees. Across conditions, the detection rate for pre-knowledge is higher than that for answer copying. This study shows the insensitivity of \( I^* \) to IRT model misspecification, and different sensitivities for different cheating types.
Session 3: Statistical Methodology and Applications

A Brief Introduction to Markov Chain Monte Carlo
Christian Meyer

Inference in Bayesian data analysis is derived from the distribution of parameter values given an observed sample. This is in contrast to analysis conducted in a frequentist regime, which bases inferences on the theoretical distribution of a statistic given repeated sampling. The parameter distribution in Bayesian statistics, called the posterior distribution, is generally quite complex and commonly does not have a closed form. Furthermore, the Minimum Mean Squared Error estimator for a parameter in a Bayesian analysis is the expected value over the posterior distribution, which requires integration with respect to the distribution function. As such, a method for efficiently approximating this integral is vital to Bayesian data analysis. With the advent of sufficient computational power, Markov Chain Monte Carlo (MCMC) has become one of the most popular methods for estimation in Bayesian statistics. MCMC exploits the flexibility of Markov chains to randomly sample from the posterior distribution in order to approximate expectations with Monte Carlo integration. In this talk, the fundamentals of MCMC are introduced. A Markov chain is defined and how MCMC produces samples from a posterior distribution is discussed including an evaluation of the estimates that MCMC produces. Additionally, motivations for developing Hamiltonian Monte Carlo, and how this relates to fitting latent variable models with MCMC is discussed.

Standard Error Adjustment for Projected IRT Scores in a Fixed-Item-Parameter Linking Design
Reni Xu

Thissen, Liu, Magnus, & Quinn (2015) linked the PROMIS pediatric and adult anxiety scales using the calibrated projection—a statistical procedure that uses item response theory (IRT) to link two tests measuring related but distinct constructs. It is worth noting that Thissen et al.’s study ignored the sampling variability due to using estimated model parameters to calculate projected scores. To address this deficiency, we propose a multiple imputation (MI) approach that produces adjusted standard errors (SEs) of IRT scale scores. The MI procedure involves drawing multiple sets of values from an approximate to the sampling distribution of item parameter estimates. In the current research, latent variable (LV; i.e., pediatric and adult anxiety) means and covariances are estimated conditional on published estimates of item parameters, such that item and LV parameters are estimated not simultaneously but rather in two stages. The asymptotic covariance matrix (ACM) of item and LV parameters in the two-stage estimation setting are calculated by pseudo-maximum likelihood estimation. We conducted a simulation study to examine the impact of carry-over sampling variability under various sample size conditions. Plug-in and MI expected a posteriori (EAP) scores and their posterior standard deviations (SDs) with or without adjustment were compared.

The Influence of Poor Measurement Quality on Model Selection: An Example Using the GRIT Scale
Weimeng Wang

Multidimensional Item Response Theory (MIRT; e.g., Reckase, 2009) is often used to explore the complex structures of latent constructs. Models within MIRT, such as the bifactor model and the simple structure model, are often applied to the large-scale assessment setting as well as relatively small-scaled psychological studies. However, the small sample size often results in poor measurement quality such as negative loadings and small loadings, which may result in misleading model fit result. This phenomenon is very common in empirical data analysis. Motivated by a particular construct of grit (Duckworth & Quinn, 2009), which recently received much attention in Educational Psychology, the current study
investigates the performance of the likelihood-based model fit indices when the measurement model has
different quality especially with negative loadings and/ or small loadings. Using a targeted Monte Carlo
simulation study, we examine how well the true model is identified between the two measurement
models, the bifactor and the simple structure model.