Error Covariance Structure in a Latent Quadratic Effects Model
Kristina Cassiday, Daniel Lee, and Greg Hancock

This presentation will discuss our research evaluating the effects of non-normal indicator error terms on estimation of latent quadratic effects. Using an unconstrained modeling approach, we observed the extent to which the parameter estimates and standard errors change when an error covariance structure is not specified for the indicators of the linear effects and quadratic effects. We demonstrate the importance of an error covariance structure mathematically, based on Mao et. al's (2014) derivations and use simulation to test varying levels of non-normality in models with and without error covariance specifications. To assess the performance of the different levels of non-normality and specified or unspecified error structure, estimates of the structural coefficients $\gamma_1$ and $\gamma_2$ were compared with the true values that were used initially to generate the data. Bias, root mean-squared error (RMSE), and standard error ratios were obtained and compared across the different non-normal conditions. Bias was calculated as the difference between the mean of the estimates across 100 replications and the true parameters. The results of our simulation support the mathematical derivations, emphasizing the necessity of including an error covariance structure when the errors are non-normal.

ANCOVA versus ANOVA: A Comparison of Bias and Precision When Groups Are Not Equivalent at Baseline
Alyson Burnett

A series of simulations investigate the rationale behind the What Works Clearinghouse (WWC) baseline equivalence standards for group design studies. The WWC evaluates baseline equivalence for quasi-experimental designs and randomized controlled trials with high levels of attrition. A study with a pretest standardized effect size difference measured by a Hedges' $G$ less than or equal to .05 satisfies the requirements, a Hedges' $G$ of .05-.25 satisfies the requirements but only when a statistical adjustment is made, such as an ANCOVA, and a Hedges' $G$ greater than .25 does not satisfy baseline equivalence. The simulations evaluate these specific criteria by comparing the bias and precision from an ANCOVA and an ANOVA analysis of the data at four levels of baseline inequivalence that map to the WWC ranges and three levels of correlations between the pretest and the outcome (12 conditions total). The first simulation generates data for a linear regression model in which the treatment status and pretest predict the outcome, and compares bias and precision for an ANCOVA analysis that uses the pretest data and an ANOVA analysis that ignores the pretest data. The second simulation, which is still in progress, tests bias and precision for ANCOVA and ANOVA in a situation in which groups are different in more than one respect at baseline, but only one covariate is available for analysis (a second covariate is added to the model but is ignored for both the ANCOVA and ANOVA analysis). The presentation will summarize results available to date, conclusions and recommendations, as well as next steps for the research.
Effects of Measurement Error in Assignment Variable on Regression Discontinuity Treatment Estimates
Monica Morell and Ji Seung Yang

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 & Retto, 2009; Hulleke, 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. 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 bandwidths are used.

Comparing Goodness-of-fit Indices Based on Ungrouped Data in Logistic Regression: A Simulation Study
Jinwang Zou

This study compared the performance of three goodness-of-fit indices in detecting omitted quadratic term in Logistic regression. The new indices differ from classic ones in that they do not require aggregating the data. A simulation study was conducted to explore the Type I error rate and power for these tests. The conditions manipulated are sample size and coefficient of quadratic term (Quadratic effect). The sample size was set at 100, 500, and 1,000 to represent small, medium and large sample size effect. The coefficient of quadratic effect was set at 0.01, 0.1, 0.2 and 0.4 to represent very small, small, medium and large quadratic effect. The results showed that the three new tests all performed well in controlling Type I error rate. Information Matrix has adequate power when sample size is large or quadratic effect is large. Unweighted Sum of Squares Test performed second best. It is useful only when the sample size is large or quadratic effect is large enough. The Standardized Pearson Test did not have adequate power in all conditions.

Session 2: Models to Detect Individual Differences and Group Processes

Fairness in Automated Scoring: Screening Features for Subgroup Differences
Ji An, Vincent Kieftenbeld, and Raghuveer Kanneganti

In automated essay scoring (AES), researchers have been aware that essays written by test takers from different populations can be scored differently by humans and machines. In recent years, researchers have presented a variety of methods to detect this unfairness issue; nevertheless, none
of the existent studies has come up with any statistical method to resolve the issue related to features that lead to unfairness. This study, through an empirical analysis, presents a statistical feature-screening approach that can be applied to detect and screen out the features that may cause the differences in subgroups. In order to do so, we used Anderson-Darling (AD) feature screening method. In particular, AD tests that compare the distributions of two subgroups (e.g., native English speakers and non-native English speakers) on each feature were conducted at each score point; features that demonstrated statistically significant differences between subgroups were removed; then the rest features were included to construct random forests to train the machine and get the automated scores. We then compared this random forests method based on selected features to that based on all features, with respect to measures of unfairness such as quadratic weighted kappa and standardized difference of the mean between human and machine scores across subgroups. Preliminary results show that the feature screening approach helped reduce the sub-group differences; the overall reliability of machine scores, on the other hand, was slightly decreased due to screening.

Social Network Modeling of 5th Grade Students’ Peer Relationships
Casey Archer

Social networks can be used to model student ties within classrooms. Recently, data was collected to measure the connections that students have during times of transition (5th-6th grade) in order to examine how social ties change as a result of transitioning from one school to the next. We have been using this data to create a model of the underlying social network that exists at the 5th grade level. Developing social networks can lead to information on students’ isolation, their number of connections, and the overall network density, which can provide more information to researchers aiming to understand how students develop from one grade to the next. Next steps will be to develop networks at the 6th grade level with the same participants in order to compare change in the networks over time.

Evaluate Subgroup Influence on Individual Behavior
Qiwen Zheng

In social science, social network analysis investigates the relations and connections of people or any other forms of entities. Subgroup structure is an inherent characteristic in many social networks. An individual’s behavior may be associate not only with his/her previous behavior but also the previous behavior of others in the same subgroup of a network. In this paper a network influence model is proposed to model the influence of subgroup structure on individual behavior. A statistical network model, the stochastic blockmodel (SBM) is used to identify subgroups based on the interaction pattern among individuals. Simulation studies with different conditions are conducted to evaluate the performance of the influence model, as well as the association of the goodness of model fit with SBM’s ability to correctly identify subgroups.

Aberrant Behavior Detection Based on a Multivariate Mixture Modeling Approach
Kaiwen Man

Recently, response times have been used to measure and evaluate many latent characteristics of students, and particular to this research, have been utilized to identify aberrant response behavior of examinees on educational and psychological tests. Statistical procedures based on response times provide a more flexible framework to study aberrant testing behavior of respondents than traditional IRT-based modeling. The focus of this research is to extend current response time methodology for detecting aberrant testing behavior to a multivariate mixture modeling framework, which incorporates two response time based person-fit indices. Detection of aberrant
testing behavior using this new mixture modeling technique will be evaluated vis-à-vis a Monte Carlo simulation.

Session 3: Issues Related to Modeling Clustered and/or Survey Data

Modeling Learning Growth with a Cross-Classified Multilevel IRT Model
Chen Li and Hong Jiao

Growth modeling has been of interest in many assessment programs, including both high-stakes and low-stakes tests. Growth could be modeled using different approaches. This study models growth with an Item Response Theory (IRT) based approach utilizing item response data. It investigates the impact of complex student clustering structure where students are cross-classified by two grouping variables on student growth modeling. Specifically, this study focuses on modeling student linear growth between two time points, along with school clustering effect and father’s occupation clustering effect using a cross-classified multilevel dichotomous IRT Model. To examine the performance of the proposed cross-classified multilevel IRT model for growth modeling, a simulation study is carried out. The magnitudes of the clustering effects for both the school and the father occupation grouping variables and the variance of the growth parameters are manipulated in this simulation study. The performance of the proposed model is evaluated in terms of parameter recovery under eight simulation conditions. The impact of ignoring the cross-classified clustering structure on students’ growth modeling and other model parameter estimation is also investigated by comparing the proposed model with the three-level multilevel IRT models in terms of parameter recovery and model-data fit.

Confidence Intervals for Intraclass Correlations in Multilevel Item Factor Models
Xiaying Zheng

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 four 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), and beta-distribution approximation (Demetrashvili, Wit, & van den Heuvel, 2014). These methods have not been comprehensively examined, especially when the outcome variable is latent. A simulation study is conducted to examine the four methods under various data conditions, including balance/unbalanced designs, sample sizes, and distributions of the latent scores (i.e., normal/non-normal distributions). The coverage probabilities of the true ICC are compared across the approximation methods.
Corporal Punishment: An Examination of School-level Components
Alison Preston

School discipline is a topic that is of high priority to school personnel, students, parents, community members, and other stakeholders. Many important questions stem from the issues related to school discipline and student rights, including the impact of various disciplines on student outcomes, the relations between the volume and types of student disciplines and student and school performance, patterns of discipline and student behavior, and so on. One particular facet that has often been overlooked in discipline research relates to a less discussed form of discipline: corporal punishment. This study uses data from the US Department of Education’s Office of Civil Rights to explore the institutional characteristics of schools which reported the use of corporal punishment and address whether those characteristics can be compiled into representative composite variables using Principal Component Analysis. Development and further exploration of these components may serve to inform stakeholders and future research into the prevalence of corporal punishment as a form of discipline in schools. The methodology of principal component analysis as well as use of this particular database for large-scale research are also examined.

Choice of Sampling Weight for Longitudinal Modeling of Panel Data
Daniel Lee, Jeffrey Harring, and Laura Stapleton

National longitudinal panel surveys in education like the Early Childhood Longitudinal Study Kindergarten (ECLS-K) face the problem of respondent attrition. Sets of panel weights are provided that allow the analyst to appropriately weight the observations in the study to account for such attrition. However, these weight adjustments are based typically on sampling design information and data collected in the base year; information obtained from early waves in the survey program is not typically utilized in the adjustment model. Another method to address potential bias from non-response is full information maximum likelihood, which uses responses obtained during all data collection waves. Using data from ECLS-K, it was found that there exist differences in estimates from analyses when non-response adjusted panel weights are used as compared with FIML utilizing the base year sampling weights only but with auxiliary information, suggesting that bias may exist when using panel weights and ignoring the attrition that is informative for a given outcome variable. The effects on bias of growth parameter estimates of using panel weights and FIML when all informative auxiliary variables (that define the missingness) are not utilized is shown via a simulation study in which the degree to which auxiliary information is correlated with the outcome and response status is varied.

Session 4: Polytomous Items and Dependence Issues in Measurement Models

An Overview of Structural Equation Modeling with Categorical Data. What Works?
Jordan Prendez

Structural equation modeling (SEM) is a popular modeling framework used across disciplines. As with many statistical methods, estimation procedures within the SEM framework make assumptions about the nature of the underlying data. Violating these assumptions may result in biased parameter estimates, inaccurate variance estimates, and unreliable inferences regarding model fit. However, these assumptions are often not understood or are implicitly ignored in practice. In particular, many researchers have employed normal theory maximum likelihood when
estimating model parameters from ordinal data (e.g., Likert scales, or binary correct/incorrect responses). While previous research has shown that with an increasing number of categories it may be reasonable to examine categorical data with methods that assume continuous data. This cutoff point in which data might be considered as sufficiently continuous differs across estimation methods and is contingent on several other factors. The current research aims to examine the complex relationship between several of these factors (e.g., Threshold symmetry, model size, number of indicators, underlying distribution) and the performance of different estimation procedures.

A Conditional IRT model for Directional Local Item Dependence in Multipart Items
Dandan Liao, Hong Jiao, and Robert W. Lissitz

New item types in English language arts/literacy tests have been introduced by the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium to assess the target learning outcomes, among which the Evidence-Based Selected Response (EBSR) is an example. The EBSR, also called a multipart item, consists of two related questions, which potentially introduces conditional local item dependence (LID) between two parts of the item. It raises concerns to use standard item response theory (IRT) models to calibrate multipart items. This paper proposes a conditional IRT model for directional LID in multipart items and compares the proposed method to four conventional approaches to accounting for LID in terms of parameter estimation through a simulation study. Number of multipart items and the magnitude of directional LID are manipulated in the simulation study.

Comparison of Classification Accuracy Based on Item Response Theory (IRT) and Measurement Decision Theory on Tests with Polytomous Items
Yating Zheng and Hong Jiao

Recently, measurement decision theory, which derives from Bayesian theorem, becomes 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 accuracy than the IRT models. However, these studies primarily focused on dichotomously scored items. This study explores the application of measurement decision theory into tests with polytomous items. Comparison is conducted in classification accuracy between the partial credit model and measurement decision theory at different test lengths with various sample sizes on tests with polytomous items using simulated data. Results indicate that in general measurement decision theory provides higher classification accuracy than the partial credit model. Among the three criteria of measurement decision theory (maximum likelihood, minimum probability of error, and maximum a posterior), maximum a posterior (MAP) criterion provides the highest classification accuracy. The classification accuracy of MAP for tests with very short test length and small sample sizes is even higher than that of the partial credit model for tests with long test length and large sample sizes.

A Combination of Diagnostic Classification Model and IRT Model with Testlet Effect
Manqian Liao

Diagnostic Classification Models (DCM) use a set of categorical latent variables to indicate the examinees’ mastery state of a set of skills in order to provide useful feedback for test-takers or
other stake-holders. However, it does not provide a general ability estimate that can be used for accountability purpose, which might not fulfill the needs of some high-stake tests. Neither does it take testlet effect into account, which could lead to the violation of local item independence assumption and biased diagnostic information. This study proposes a model that combines testlet IRT model with DCM with two aims. Firstly, the proposed model intends to take into account local item dependence due to the testlet effect. Secondly, it aims at realizing dual purposes (both accountability and diagnostic purposes) within one test administration. In the proposed model, mastery states, continuous trait, and testlet effect will be estimated simultaneously. The parameters will be estimated with Maximum Likelihood Estimation and a simulation study will be conducted to evaluate the estimation quality.

Session 5: Issues in Measurement of Health, Academic, and Team Constructs

Measurement of Transactive Memory Systems
Kylie Goodell King

A transactive memory systems (TMS) is a team-level construct that is comprised of distributed team knowledge and the processes by which that knowledge is shared amongst team members (Wegner, 1987). TMSs have been measured in laboratory and field settings by researchers from a number of different backgrounds, including social psychology, organizational behavior, and education. Past measurement techniques include using outcome measures to infer the existence of a TMS (for example, the ability to recall words individually and with a partner), rater evaluations of team tasks (for example, radio assembly), and self-report indicators (including a 15-item inventory that has been used by a number of scholars over the past decade). This review summarizes previously employed measurement techniques and discusses alternative approaches to TMS measurement, including collecting and using multilevel data.

Measures of Health Disparity in Complex Samples
Meng Qiu and Yang Li

Health disparities refer to differences in the health status across different racial, ethnic, and socioeconomic groups. Some groups of people have higher rates of certain diseases, and more deaths, and suffering from them, compared with other groups. Great effort has been made to identify and address the factors that lead to health disparities and many summary indices (e.g., Concentration Index) have been developed to measure the extent of health disparities. Traditional indices assume that data were obtained through simple random sampling (SRS), with each individual having an equal probability of being included in the sample. For large, nationally representative health surveys, however, complex sampling designs are often employed to select the sample, involving stratification, multistage clustering, or both. This study proposed a new index, based on Absolute Concentration Index (ACI) which incorporates the complex sample design features to measure health disparities using data from nationally representative samples with complex designs. The variance of the proposed measure was estimated by Taylor linearization method. Simulation study and real data analysis will be reported if available.
Small-Sample MIRT Calibration and Model Selection using the Grit Data
Wanyi Liu, Ji Seung Yang, Allan Wigfield, and Katherine Muenks

Recent developments in multidimensional item response theory (MIRT; e.g., Reckase, 2009) provide more options in choosing flexible item response models that reflect complex structures of psychological constructs. Accordingly, MIRT applications have become more common not only for large-scale assessment data (e.g., Programme for International Student Assessment or Patient Reported Outcomes Measurement Information System) but also for relatively limited data collected in research settings (e.g., Reise and Waller, 2009). However, the small sample size issue in MIRT can easily puzzle researchers with respect to convergence of the model solutions and model fit diagnostics because the typical asymptotic properties are not guaranteed. The purpose of this study is to illustrate the phenomena and demonstrate utility of a targeted simulation study (known as parametric bootstrap) in examining the structure of grit, a newly emerging construct that has gained substantial attention recently in education and psychology (Duckworth and Quinn, 2009). With the empirical data, using different model fit indices leads to divergent conclusions. A parametric bootstrap simulation helps explain the behaviors of likelihood-based model fit indices, and provides guidance to choose the best statistic when different statistics disagree. Preliminary results found that the power of Bayesian information criterion (BIC) to detect a true model is very low with a small sample size when the data generated from a bifactor model, while Akaike information criterion (AIC) is more efficient with small sample size to detect a true data generating model.

The Healthy Eating Score: A Five-Item Measure of Diet Quality
Golenbock, S., Kazman, J. B., Moylan, E., Kupchak, B., and Deuster, P.

Dietary assessments are time-consuming. The present study sought to validate a 5-item Healthy Eating Scale (HES-5) against a comprehensive food frequency questionnaire (FFQ) that yielded an overall Healthy Eating Index (HEI; as recommended by the USDA); nutritional biomarkers (Omega-3, lipoproteins, fasting glucose); and other health variables (healthy eating practices, activity levels, body fat, Army Physical Fitness Test scores). 221 soldiers (77% male; aged 28±9), from four Army installations, completed the HES-5 (which asks about fruit, vegetable, dairy, whole grain, and fish consumption), a FFQ (based on 110 foods), additional questions and laboratory measures. The HES-5 strongly correlated with HEI (r=.48). Among the biomarkers, the HES-5 and HEI only correlated with Omega-3. Both were moderately related to percentage body fat, and to other healthy eating practices (eating breakfast and recovery snacks after exercise). Only the HES-5 correlated with activity levels, and neither score correlated with reported Army Physical Fitness Tests. The HES-5 score appears to be an adequate indicator of a healthy diet when compared with scores for the USDA's HEI among soldiers. The HES-5 should be considered as a brief measure of diet quality when resources are not available to administer a full FFQ.