IMPS 2016

Asheville, NC, USA • July 11-15, 2016

Abstract Book: Posters

Tuesday, July 12, 2016

Poster Session: 6:00 PM - 8:00 PM

(APP) APPLICATIONS

Poster 1: Measuring Grit Among First Generation College Students: A Psychometric Analysis

Brooke Midkiff, The University of North Carolina at Chapel Hill; Michelle Langer, The University of North Carolina at Chapel Hill; James Ellis, The University of North Carolina at Chapel Hill; Cynthia Demetriou, The University of North Carolina at Chapel Hill; Abigail Panter, The University of North Carolina at Chapel Hill

The concept of grit is of interest in the field of education, particularly as it pertains to persistence to a 4-year college degree. This study offers an IRT analysis of the Grit Scale (Angela Lee Duckworth & Quinn, 2009; Angela L Duckworth, Peterson, Matthews, & Kelly, 2007) when used amongst first generation college students (FGCS) as well as recent first generation college graduates and non-FGCS recent graduates. The Grit scale was included in surveys administered as part an array of other research projects within The Finish Line Project – a U.S. Department of Education First in the World grant funded project that seeks to improve FGCS access to, persistence in, and completion of postsecondary education, through rigorous research into various programs and supports for FGCS's. A group of 155 FGCS's currently enrolled completed the original 12-item Grit scale, 181 recent graduates who were FGCS's completed a 9-item version of the scale, and 253 non-FGCS recent graduates completed the same 9-item scale. Reliability analysis, factor analysis, item response theory, and differential item functioning were used to analyze both versions of the Grit scale. The reliability and validity of the Grit scale has not yet been analyzed for use with first generation college students, or with general research-1 university students. By comparing enrolled students and graduates, the psychometric analysis in this study offers insight into the measurement of student grit for use in support program development and policy-making to improve student retention.

Poster 2: Alfred Binet, Lewis Terman, and Modern Intelligence Testing

Valerie Ryan, The University of Rhode Island

The creation of the first modern intelligence tests was significant because it combined the use of statistical analyses and experimental methodologies with the study of complex mental processes, which was not a common practice at the time. The two psychologists most responsible for the generation of the tests were Alfred Binet and Lewis Terman, who created the tests within the socio-cultural context of early twentieth century France and the United States. The creation and dispersion of the tests illustrates the recursive relationship of psychologists as exerting substantial influence on society, while in turn being influenced by cultural issues of the day. Binet believed that intelligence is largely influenced by environmental factors and can be partially improved through training. His work focused on improving education for children in public schools in Paris, particularly for children with intellectual disabilities. Terman's main work included a major revision of Binet's intelligence scale, a longitudinal study of gifted children, and the administration of intelligence tests for military recruits during World War I. Unlike Binet, Terman believed that intelligence is hereditary and linked his work to eugenic theory. These intelligence tests lead to an expansion in the field of mental testing, resulting in the creation of many tests that are still used in psychology today. The tests also had proximal societal impacts in the United States: mental testing led to stricter immigration laws, an increase in eugenic practices, and major changes in the educational system.

Poster 3: WITHDRAWN

Poster 4: WITHDRAWN

Poster 5: An Proposal Measure the Relationship Walking and Body Composition Index
Kotato Ohashi, Rikkyo Unoversity; Yuko Oguma, Keio University Sport Medicine Research Center; Michiko Watanabe, Keio University

In this paper we proposed a statistical analysis method to apply the growth curve model. Also we showed an example of analysis based on biometric log data obtained from the accelerometer and body composition meter. Specifically, we decided the number of groups of subjects by using the Growth Mixture Model to their cumulative number of steps per week at first. The next, we applied Analysis of Covariance to this result, and confirmed whether differences in body composition were observed between the groups or not. Then, we found that the factor of grouping that is cumulative number of steps correlates visceral fat.

Poster 6: The Evolution of Psychometrics: Genealogical Trees of Psychometric Society Presidents
Lisa Wijsen, University of Amsterdam; Willem Heiser, Leiden University; Denny Borsboom, University of Amsterdam

Psychometrics as a discipline evolved in the nineteenth century as the joint product of psychology's attempt to measure psychological traits and states and society's pressure to develop objective instruments for educational testing (Porter, 1995). Even though psychometrics has played such a central role in our understanding of human beings, relatively little research has been done on its history. How psychometrics has developed, and how this development has changed our conception of measuring the human mind, are both areas of research that are generally overlooked. An important factor in the evolution of a scientific field is how knowledge and skills are transferred across generations of scientists. In this talk I will present some genealogical trees of the major scholars in the history of psychometrics. These genealogical trees show the supervisor-student relations of the current and past presidents of the Psychometric Society, tracing back to the beginnings of psychometrics. A genealogical tree is a useful tool to visualize the development of a scientific field, and provides us with interesting hypotheses regarding the historical and current state of psychometrics. Are all psychometricians part of the same tradition, or is psychometrics highly fragmented? Is psychometrics an independent scientific field or is it strongly connected to other fields, such as mathematics, econometrics or statistics? In this talk, I will present the genealogical tree and focus on what this tree implies for the evolution of psychometrics.

Poster 7: Change Your Attitude! Attitudinal Survey Characteristics Impacting Participant Responses
Elisabeth Pyburn, James Madison University; Deborah Bandalos, James Madison University

Our research focuses on manipulating features of affective items hypothesized to affect response processes. This research is similar to Embretson’s (1983: 1998) work with cognitive test items. We present the results of preliminary studies in which we manipulated item characteristics experimentally to determine their impact on item intensity. As one example, research has indicated that vague quantifiers (“many” or “seldom”) can have different meanings for different people (Wright, Gaskell, & O’Muircheartaigh, 1994). We assessed the impact of four item characteristics on participant responses: (a) vague wording, (b) question vs. statement phrasing, (c) labeling vs. not labeling center response options, and (d) using unbalanced response options (e.g., more Agree than Disagree options). We created two versions of each questionnaire: a control version (original wording) and a manipulated version in which the four item characteristics were varied. The versions were combined to create four different forms, each of which was completed by approximately 125 undergraduate students. We also conducted think-aloud interviews with eight students to determine how the manipulations impacted their responses to the items.

Examination of group mean differences, reliability statistics, and item parameters via confirmatory factor analysis invariance analyses revealed some interesting differences among manipulations. Results from the think-aloud interviews helped further explain these findings. Based on the results from this first study, we chose two manipulations – vague wording and question/response format – for further study. We administered these two manipulations to a larger group of students (n ≈ 2000), and are currently analyzing the results.

Poster 8: The Relationship of Decision Regret and Patient Factors

Kiyomi Tanno, Rikkyo University

Purpose

During the treatment process, it is essential that the patient feels satisfied with the decisions made to proceed with treatment. In order to evaluate the level of satisfaction of decision-making, it is extremely useful to assess the quality of treatment based on the concept of regret.

The principal aims of this study were to elucidate the relationship between regret and patient factors in the Japanese DRS.

Methods

We administered a questionnaire survey concerning benign uterine and ovarian tumors that received operative treatment. We then performed hypothesis testing of the relationship between the Japanese DRS, health-related QOL, and patient factors using latent class analysis and path analysis through a multipopulation comparison.

Results

A total of 102 patients were targeted for study. Through the latent class analysis of patient characteristics, patients were classified into two groups: patients who were married and had children, and patients who were unmarried and had no children. The path analysis through multipopulation comparison of two classes revealed that a subjective symptoms, preference, or surgical procedure 2 [laparotomy or laparoscopy] had a direct impact on regret.

Conclusion

The present study examined the relationship between regret and patient factors. It was revealed that patient factors having direct impact on regret. Study results suggest that in the case of benign tumors, patient behavioral factors influence decision-making regret during the treatment process. It is necessary for physicians to support decision-making while considering the hopes and fears of the patient based on characteristics of the disease.

Poster 9: A Mixed-Methods Approach to Differential Item Functioning

Madison A. Holzman, James Madison University; S. Jeanne Horst, James Madison University

As diversity within higher education increases, it is important to ensure test scores are psychometrically sound and unbiased for all groups of students. Historically, international students at a predominately white public university score lower than domestic peers on a university-required information literacy test. This test consists of 54 common items and 36 additional items spread evenly across three forms. We evaluated international (N=87) and domestic students' (N=4,383) scores for differential item functioning (DIF) favoring domestic students. Three items were flagged for DIF, via both Mantel-Haenszel and logistic regression techniques. To further evaluate why the items displayed DIF, international (N=13) and domestic students (N=16) completed think aloud protocols (TAPs). The TAPs methodology in the current study differed slightly from typical verbal concurrent TAPs. Following a short training session, rather than verbally expressing their thoughts, participants wrote their thoughts to each question. Throughout the TAPs, participants were reminded to write their thought processes as they answered each item. Employing a mixed-methods explanatory-sequential design, qualitative data from the TAPs were combined with quantitative data from the DIF analyses to provide a framework for why items may be functioning differentially. We consulted with content experts during each phase of the research study. Limitations include discrepancy in sample size between the two student groups and amount of information provided by students during the written TAPs. Though written TAPs presumably yielded less information than verbal TAPs, we believe there could be a place for written TAPs when sample size and resources are of concern.

Poster 10: Do Standard Setting Panelists Really Understand Angoff Method?

Hong Qian, National Council of State Boards of Nursing; Mark Reckase, Michigan State University; Doyoung Kim, National Council of State Boards of Nursing; Ada Woo, National Council of State Boards of Nursing

Standard setting panelists are subject matter experts for the examination, but usually do not have knowledge of standard setting methods before attending standard setting workshops. The understanding of standard setting method is critical for panelists to perform their tasks confidently and accurately during the workshop. However, currently it is typically measured by self-reported evaluation questions using Likert scales. This study develops an objective assessment to measure panelists' knowledge of a commonly-used standard setting method—Angoff method. This assessment can provide better information about whether panelists really understand the method.

The assessment consists of 12 questions and has been developed by a faculty member teaching standard setting and several testing practitioners who have conducted standard setting workshops repeatedly. This assessment has been administered to two separate standard setting panels (5 and 10 panelists respectively) at the end of standard setting workshops along with the self-reported evaluations. With the help of this assessment, several interesting research questions can be explored: (1) How is the overall understanding of Angoff method among the panelists? (2) Are there any associations between assessment performance and self-reported evaluations? (3) Are there any associations between assessment performance and standard setting ratings? (4) If the ratings from the low performer are excluded, will the reliability of ratings increase? This assessment will be administered to another standard setting panel in July 2016 thus more data will be available for addressing the research questions. The results from this study have significant implications for standard setting training and setting appropriate passing standard.

Poster 11: Comparison of Pre-Equating and Post-Equating Across Different Linking Conditions

Hyeonjoo Oh, Educational Testing Service; Junhui Liu, Educational Testing Service; Hanwook Yoo, Educational Testing Service

Most equatings conducted for many testing programs are post-equated, however, some testing programs have to estimate an equating function before a new form is operationally administered to expedite score reporting. Despite this appealing advantage, findings from the earlier studies on pre-equating are controversial and inconclusive. Given the continuous use of pre-equating operationally for many testing programs despite the inconsistent results, there are still other factors that may affect the quality of pre-equating, such as the characteristics of the anchor test. This study compares pre- and post-equating with three different anchor conditions (i.e., hard, similarly difficult, and easy anchors), two different group ability differences (i.e., moderate and small ability differences), and five linking methods (i.e., Mean/Mean, Mean/Sigma, Haebara, Stocking-Lord, and fixed parameter calibration) using large scale assessment data. For the current study, Rasch model was used for calibrating and linking. The pre- and post-equating designs are evaluated in comparison with a criterion equating function using average absolute deviation and root expected squared difference. The preliminary results of the current study suggest that when similarly difficult anchor is used and the ability difference between the new and old form groups is small in post-equating, both pre- and post-equating showed similar equating results. The findings may help practitioners to improve the performance of pre-equating in various equating situations and help to ensure the quality and fairness of assessment results.

Poster 12: Small Sample Linear Equating for Anchor Test Design

Nurliyana Bukhari, The University of North Carolina at Greensboro

Most of the equating studies have been done based on samples of several thousand examinees. In certain testing situations, this condition is often implausible. Certification and licensure tests for instance involve very small target populations but often administered more than a single administration annually. Therefore, equating with small sample of examinees is more challenging than equating using large samples (Sunnassee, 2011; Parshall, Houghton, & Kromrey, 1995).

The current simulation study was conducted to investigate the bootstrap standard errors and statistical bias of equating for the non-equivalent anchor test (NEAT) design using the linear equating technique with small sample examinees. The preliminary results revealed that despite the large equating errors, the level of equating bias seems inconsequential. However, equating under the NEAT design with sample size of 50 or smaller must be conducted carefully due to the inconsistent results and the large standard error especially with sample of 25 examinees. The Tucker method tends to work best, given the criteria and limitations of the study. The Braun-Holland method tends to work the least given that it adopts the frequency estimation assumptions to conduct linear equating (Kolen & Brennan, 2004). Frequency estimation that is derived from the equipercentile function is more susceptible to sampling error than the linear function because it involves the estimation of as many parameters as there are unique score points on the original form of the test (i.e., Form X) (Albano, 2014).

Poster 13: A Comparison of Item Parameter Recovery Across Different R-Packages

Taeyoung Kim, State University of New York at Buffalo; Insu Paek, Florida State University

With the advent of the free statistical language R, several item response theory (IRT) programs have been introduced as psychometric packages in R, which have an advantage of a free open-source over commercial software. For research and possibly considering the use of these free R programs, one may wonder the quality of the results produced by these free programs. The goal of this study is to provide information regarding the performance of those free R IRT software for item parameters and their standard error estimation. The study conducts a series of comparisons via simulations, for the Rasch, 2-parameter logistic (2PL), and 3-parameter logistic (3PL) model. The IRT programs which are rigorously evaluated in the present study include the most updated versions of “eRm” (Mair et al., 2015), “ltm” (Rizopoulos, 2013), “mirt” (Chalmers, 2016), “sirt” (Robitzsch, 2015), and “TAM” (Kiefer et al., 2015). The authors employ different conditions concerning the number of examinees and items to reflect various situations in practice. Also this study reports which standard error methods are used for each package, whether the convergence is achieved at each replication for both “eRm” and “ltm”, and the elapsed times for the estimation of the models in the different simulation conditions. Our preliminary results suggest that bias and root mean squared errors (RMSE) of item parameter estimates and their standard errors in “eRm”, and “ltm” for the Rasch model were nearly the same, while those in “ltm”, “TAM”, “sirt”, and “mirt” differed noticeably for the 2PL model.

Poster 14: 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.

Poster 15: Habitual Physical Activities and Academic Achievement in Basic Education

Tong Ran, Beijing Normal University; Jichao Zhong, Beijing Normal University; Danhui Zhang, Beijing Normal University

For the past decades, the effects of physical activity (PA) on student academic achievement were wildly investigated from multiple perspectives. However, empirical studies on this subject were scarce in China. This study employed Hierarchical Linear Models (HLM) and Structural Equation Model (SEM) to address the following two questions: 1) To what extent PA will influence students' academic performance in mathematics, while considering for social-economic factors, gender and grade; 2) How learning motivation and self-efficacy mediate the relationship between PA and students' academic performance in mathematics, based on the secondary data analysis of a large scale assessment in China. It was found that there was positive effect of habitual PA towards academic performance. Yet, such influences were not consistent regarding with different groups, e.g. lower grade or male students. Furthermore, students from economically advanced background had better development in habitual PA, as well as a larger effect size in the regression model. Moreover, learning motivation and self-confidence were discovered to act as significant mediators. Better PA habits positively related with higher motivation, thus in turn leading to better academic achievement.

Poster 16: Pre-Service Teachers' Attitudes Toward Inclusive Education for Exceptional Children

Huang Chiu-Hsia, National Pingtung University

The investigation aims to develop the Inclusive Education Scale divided into three subscales, Concepts of Inclusive Education, Expectations of Social Justice and Perceptions of Disability Groups Scale, based on Torres-Harding, Siers and Olson (2012); McHatton and McCray's (2007) theories. Firstly, it aims to investigate 62 sophomore and senior pre-service teachers' supportive attitudes toward inclusive education for exceptional children. Secondly, all items of Inclusive Education Scale are considered on the multicollinearity and regression model, all data analyses are descriptive by the frequency, percentile, mean, SD, $p<=0.05$, Cronbach's α, and item analysis.

Overall, the Cronbach's α of the Concepts of Inclusive Education, Expectations of Social Justice and Perceptions of Disability Groups Subscales separately are 0.698, 0.905 and 0.884 respectively; it means that this entire set with three subscales is a very stable and reliable questionnaire. In addition, the senior pre-service teachers are more willing to support exceptional children than sophomore pre-service teachers do.

The investigator will advise all teacher training institutions providing all pre-service teachers more volunteering experiences to contact with exceptional children. Taiwan government may emerge the inclusion issues into K-12 even continuing education programs in order to educate all Taiwanese civil with more supportive attitudes toward those who with special needs in their daily lives.

Poster 17: Various Pre-Service Teachers' Attitudes Toward Inclusive Education for Exceptional Children

Huang Chiu-Hsia, National Pingtung University

The investigation mainly aims to compare various majoring in special education, early childhood and education pre-service teachers' attitudes toward exceptional children by the Inclusive Education Scale divided into three subscales, Concepts of Inclusive Education, Expectations of Social Justice & Perceptions of Disability Groups, based on Torres-Harding, Siers and Olson (2012); McHatton and McCray's (2007) theories. All items of Inclusive Education Scale are considered on the multicollinearity and regression model, all data analyses are descriptive by the frequency, percentile, mean, SD, p<=0.05, Skewness, Kurtosis, Cronbach's α, item analysis, factor analysis and correlation.

191 pre-service teachers include of 55 (28.2%) majoring in education, of 61 (31.3%) majoring in special education and of 75 (38.5%) majoring in early childhood those who participate in this investigation. Overall, the Cronbach's α of the Concepts of Inclusive Education Scale, Expectations of Social Justice Scale and Perceptions of Disability Groups Scale separately are 0.639, 0.924 and 0.877 respectively; it means that this entire set with three subscales is a very stable and reliable questionnaire. In addition, majoring in the special education pre-service teachers are more willing to support exceptional children than majoring in early childhood or education pre-service teachers do; majoring in early childhood pre-service teachers are more willing to support exceptional children than majoring in education pre-service teachers do.

(BSI) BAYESIAN STATISTICAL INFERENCE

Poster 18: A Bayesian Multilevel Modeling Analysis of PISA with Informative Priors

Jing Yuan, University of Kentucky; Hongwei Yang, National Board of Osteopathic Medical Examiners

This paper aims to understand U.S. middle school students' mathematics achievements by examining relevant student- and school-level predictors. To that end, we derive two datasets from the Program for International Student Assessment (PISA) with the primary one from 2012 (4,978 students) and the secondary one from 2003 (5,137 students). The dependent variable (DV) is a composite measure of mathematics literacy, calculated from an exploratory factor analysis of all five PISA mathematics achievement plausible values in each dataset for which evidences are found supporting data unidimensionality. Through a variance component analysis of each DV, the use of multilevel modeling is warranted for analyzing each dataset. Then, a multilevel analysis of the primary dataset is performed under the Bayesian framework with Gibbs sampling and conjugate priors for all structural and covariance parameters. Simultaneously, the secondary dataset is analyzed under the traditional maximum likelihood method to provide the information needed to specify informative priors for the Bayesian analysis of the 2012 dataset. The cycles of PISA are known to be implemented and administered under comparable conditions so that the use of results from the prior cycles as informative priors is justified. During the analysis, predictors are entered sequentially in the order of theoretical importance to create a hierarchy of models. By evaluating each model using Bayesian fit indices, a best-fit and most parsimonious model is selected for interpretation and discussion. The predictors include demographic and content-specific variables: mathematics efficacy, teacher-student ratio, etc. Finally, the analysis is performed using R MCMCpack and MCMCglmm packages.

Poster 19: Understanding Marriage Duration in the US Using Bayesian Survival Analysis
Qun Zhang, University of Kentucky; Hongwei Yang, National Board of Osteopathic Medical Examiners

This paper demonstrates an application of tackling censored data in divorce counseling. Censoring is a typical problem in therapeutic data analysis—some couples had not yet divorced by the time of last observation, resulting in the event of interest (i.e., marriage duration) being only partially known. To that end, a survival analysis featuring Gaussian censored regression is performed under SAS to predict the time to the occurrence of divorce.

Using data from the National Survey of Families and Households, predictors including couples' marital conflict resolving strategies, perceived fairness in domestic tasks, time spent together and additional couples' characteristics (e.g., religion differences, sexual satisfaction, etc.) are analyzed under a Bayesian framework with Markov chain Monte Carlo (MCMC) sampling. Conjugate priors for the likelihood function from the Gaussian censored model are specified on all model parameters to factor in prior empirical evidence in family sciences. Predictors are added sequentially in a hierarchical manner to build multiple lifetime models, which are evaluated for convergence and goodness-of-fit using relevant Bayesian statistics and plots. Lastly, Bayesian statistical inference is conducted on the finally selected model to describe the relationship between marriage duration and the predictors.

SAS PROC LIFEREG, featuring parametric failure time models in Bayesian analysis, along with Gibbs sampling MCMC algorithm for simulating all posterior samples of model parameters, may contribute to a deeper and more intuitive understanding of the research question. The full prediction of marriage duration distribution obtained through Gaussian processes makes censored survival data more clinically informative.

Poster 20: Stochastic Model Comparison Methods with Highly Correlated Latent Traits
William Muntean, Pearson; Joe Betts, Pearson; Jing-Ru Xu, Pearson; Ada Woo, National Council of State Boards of Nursing

Many cognitive based latent traits are highly correlated. For example, clinical decision-making ability may be highly correlated with clinical knowledge and skills. Although highly correlated, the theoretical relationship is distinct and important. However, when sets of observations are highly correlated, factor analytic solutions can have difficulty separating them. Parameters close to their boundaries, e.g. very high correlations, pose problems for estimation techniques that maximize a likelihood. This propagates to model comparison indices, and therefore, psychologists should be cautious when using maximum likelihood (ML) estimation to determine the factor structure of highly correlated latent traits. Stochastic estimation techniques, such as Markov chain Monte Carlo (MCMC), are often less susceptible to boundary issues, especially when constraining prior distributions. However, model comparison techniques under stochastic estimation are cumbersome and analytically intensive, preventing their wide use. For example, calculating proper Bayes Factors require special sampling algorithms (e.g., reversible jump MCMC, transdimensional MCMC, product-space MCMC) that are increasingly complex as model dimensionality increases. Fortunately, pseudo Bayes Factors are easily calculated from summing across the log of the conditional predictive ordinate, which are readily accessible from MCMC samples. The current study investigates two estimation techniques using two-dimensional data with highly correlated factors and evaluates the extent to which the strong correlation contaminates model comparison.

Poster 21: Bayesian Model Averaging for Multilevel Models

Yuzhu Yang, University of California-Merced; Sarah Depaoli, University of California-Merced; Keke Lai, University of California-Merced

Model selection can be challenging when the relationship between the outcome and a set of predictors is unknown to the investigator. Bayesian model averaging (BMA) may help with the decision of which predictors to include in the model. BMA does not select a single "best" model. Rather, it assumes a list of possible models exists. BMA computes the posterior of each possible model and uses it as the weight to average the posteriors for the parameters across all models. This process produces an averaged posterior distribution of the parameter to be estimated. Then the posterior mean and variance can be computed to represent the final estimate of this parameter. BMA is widely applied in a variety of areas within statistics but has yet to be extending into multilevel modeling (MLM) in the literature. In this paper, we combined the BMA technique with MLM. We evaluated the performance of model averaging over different multilevel models within the Bayesian framework using a simulation study. In this simulation study, we examined a two-level random intercept and random slope model with two level-1 (individual) predictors and two level-2 (group) predictors. We also compared the BMA approach with a Bayesian MLM (without the averaging component) in terms of parameter recovery. Next, we provided a tutorial on conducting a BMA for MLM with an empirical example. We concluded with a discussion of the benefits and challenges for implementing BMA in the context of MLM.

(CBT) COMPUTER-BASED TESTING

Poster 22: Effect of Item Position and Preknowledge in Computerized Adaptive Testing

Alex Brodersen, University of Notre Dame

Within high-stakes testing programs there is a constant concern of compromised or "stolen" items. This concern has been met with a growing body of literature on the detection and subsequent treatment of these cases. However, there is little investigation into the extent to which item preknowledge affects ability estimation. A purported advantage of computerized adaptive testing (CAT) is the large size of the item bank in comparison to the actual number of administered items, making it less vulnerable to item theft. A second advantage of CAT is its self-adjusting nature which may reduce impact of preknowledge. However, in reality item sharing and pooling (Chang & Zhang, 2002) is a serious issue for CAT. Additionally, as has been noted by Chang & Ying (2008) and Cheng & Liu (2015), step size in CAT can be much larger earlier in a test than later. Consequently, if an examinee has preknowledge on an item that appears early in the test, it can have a large effect on the ultimate ability estimate. Furthermore, under some common item selection procedures, highly discriminating items are selected first, leading to large step size and producing a compounding effect. We first present an analytical derivation of the expected gain in the ability estimate when an examinee gained pre-knowledge on an item. Then, we present the results of a simulation study conducted to investigate the gain in ability estimate with item pre-knowledge across various conditions including varying test length, item selection procedures, ability levels, and location of compromised items.

Poster 23: Investigating Constraint-Weighted Item Selection Procedures in Unfolding CAT

Ya-Hui Su, National Chung Cheng University

Since examinees are given different sets of items from a large item bank, computerized adaptive testing (CAT) not only enables efficient and precise ability estimation but also increases security of testing materials. The construction of assessments usually involves fulfilling a large number of non-statistical constraints, such as content balancing, key balancing, item exposure control, etc. To improve measurement precision, test security, and test validity, the maximum priority index approach can be used to monitor many constraints simultaneously and efficiently in CAT. Many previous CAT studies were investigated for dominance items;

however, only few CAT studies were investigated for unfolding items. In practice, an attitude measurement or personality test, such as the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) or Cattell's 16 Personality Factors Test (16PF), might fit better with the unfolding models than with the dominance ones. Besides, these tests commonly have hundreds of items from complex structures. Therefore, the purpose of this study is to investigate constraint-weighted item selection procedures in unfolding CAT.

Poster 24: Rescoring Methods for Flawed Items in a Scripted CAT

Chunxin (Ann) Wang, ACT, Inc.; Jie Ji, ACT, Inc.; Yi He, ACT, Inc.; Lisa Gawlick, ACT, Inc.; Nancy Petersen, ACT, Inc.

Despite every effort test developers make to guarantee item soundness, a flawed item that was intended for scoring does occur occasionally in administration. When a flawed item occurs in linear paper and pencil tests (P&P), rescoring methods commonly employed are either removing the item from scoring or rescoring the item as correct. To investigate whether these rescoring methods are still applicable to the computer adaptive test (CAT), Potenza and Stocking (1997) conducted a simulation study finding that these methods still worked well with the CAT. The simulation study in Potenza and Stocking (1997) used the weighted deviation model as the item selection algorithm in the test. However, an item selection algorithm based on a probabilistic model such as the weighted deviation model does not guarantee every test administered meets the test specifications. Lee, Li, Petersen and Gawlick (2014) introduced Scripted Testing (McKinley, Petersen & Spray, U.S. Patent No. 8,834,173 B2., 2014) which overcomes shortcomings of the traditional CAT. Whether the conclusions based on the Potenza and Stocking (1997) study could be generalized to CAT tests that employ different item selection algorithms is worthy of further evaluation. Therefore, the purpose of this study is to investigate rescoring methods in a CAT under Scripted Testing and to evaluate whether the item parameters and the administration positions of the flawed items would have any impact on the rescored results. The study will provide information on which rescoring method is more preferable given the administration positions and the item parameters of the flawed items.

Poster 25: A SAS Macro for Optimal Test Assembly

Jia Ma, The University of North Carolina at Greensboro; Shuying Sha, The University of North Carolina at Greensboro

Among optimization heuristic methods, the normalized weighted absolute deviation heuristic (NWADH) has gained popularity in automatic test assembly (ATA) for solving large and complex problems (Luecht, 1998). NWADH is a greedy algorithm (Nemhauser & Wolsey, 1988) which keeps local searches to admit one item or item set at each step for considering all the test specifications in the object function and offers a near-optimal solution on matching item information function (IIF) to target test information (TIF) in item response theory (IRT). Compared with linear programs, NWADH shows advantages of: 1) incorporate with large numbers of constrains; 2) always yield a solution with possible minor constrains violations; 3) quick computer processing time; 4) the only heuristic has been used for multistage test assembly (Zheng, 2012). However, the procedure of NWADH is not published in any software package. As SAS is one of the most popular statistical software for psychometric research, a SAS macro is written by SAS syntax as a user-written package. The object of this presentation is to discuss the use of a SAS macro implementation of NWADH. An example is provided for demonstrating this SAS macro. The specification parameters, options, data inputs, and output is discussed in the presentation. This macro provides a useful tool for researches who wish to use NWADH for ATA in psychometric research with SAS software.

Poster 26: Computer Classification Testing: Effective Screening of Identifying Youth Mental Health

Jihye Kim, Kennesaw State University

Mental health screening aims at detecting children at behavioral emotional risk and providing information if children need any mental health service or special care. The Behavioral and Emotional Screening System (BESS; Kamphaus & Reynolds, 2007) is one of the screening practice. The format of the BESS, however, is pencil-paper type and the computerized format has not been implemented yet. Currently, computer classification test (CCT), has been increasingly adopted in testing and research area (Smits, Finkelman, & Kelderman, 2016; Barrett, 2015; Fliege et al, 2005; Simms & Clark, 2005; Triantafillou, Georgiadou, & Economides, 2007). The benefit of the CCT is that test can be administered different test length to individual and classify examinees into certain categories, providing reducing testing administration time. The purpose of this study is to examine the BESS how many items can be administered with capturing the information of students' behavioral and emotional risk. In particular, of three BESS Forms (student, teacher, & parent), the teacher form is of interested. Teachers should evaluate and rate multiple times for their students with the full length of test (N=27 for each student), which causes to aggravate teachers' additional workloads. We analyzed the data (N=944) at elementary schools in the Los Angles United School District (LAUSD), using a post-hoc simulation with a termination criteria and found that the most teachers needed shorter items (less than 5 items).Thus, CCT on BESS was proven to be an efficient testing for some distinct stakeholders, such as teachers.

Poster 27: Improving the Efficiency of Exposure Control in Computerized Adaptive Testing

Shu-Ying Chen, National Chung-Cheng University; Huang Chiu-Hsia, National Pingtung University

Even though computerized adaptive tests (CATs) have become popular in the field, the threat of item sharing remains. To reduce the threat of item sharing in CATs, the Sympson and Hetter procedure with general test overlap control (SHGT) was proposed. In the SHGT procedure, the item exposure rate is controlled by adopting the SH procedure, while the general test overlap is controlled by implementing the GT procedure. By combining these two procedures, the two indices can be simultaneously controlled. The SHGT procedure, however, does not perform efficiently in exposure control. Although item exposure rate and general test overlap rate are defined differently, they are closely related and can not be controlled independently. When items are administered to all examinees, the item exposure rates would be as high as 1.0, and the general test overlap would not be low. On the other hand, when items are administered with strict control on item exposure rates, the general test overlap would not be high. Thus, the general test overlap would be affected not only by the GT procedure but also by the SH procedure. For the item exposure rates, similar situation is applied. Controlling these two related indices by implementing two separate procedures might not be an efficient approach. To improve the efficiency of exposure control, the relationship between the two indices should be considered. The purpose of this study is to propose a procedure to improve the efficiency of exposure control by taking the relationship between the two indices into account.

Poster 28: The Impact of Local Item Dependence on Mixed-Format MST

Yuxi Qiu, University of Florida; Sari Halil, University of Florida

The use of mixed-item format tests that consist of both dichotomous (e.g., true/false) and polytomous (e.g., open-ended) in large scale testing programs is increasing. This is because in terms of validity and reliability, mixed-item format tests combines the advantages of dichotomously and polytomously scored tests. One of the important issues with mixed-format of multistage testing (MST) is to eliminate dependence between items, which is ensuring items in a test are not related to each other. This is because local item indepencence is the cornerstone of item response theory (IRT), and violation of this assumption (i.e. LID) may result in negative consequences on the estimate of student ability. The purpose of this study is to explore the impact of violating local item independence on mixed-format MST. To systematically investigate the performance of mixed-format MST, a simulation study is designed in which there are a variety of manipulated simulation factors including, MST panel structure (1-3 vs. 1-3-3), test length (30 vs. 60-item), proportion of polytomous items (0%, 25%, 50% vs. 75%), and proportion of items violated local independency (5%, 10%, 15%, vs. 20%). The 3PL and partial credit model is used for dichotomous and polytomous items, respectively. 5000 examinees are generated from N(0,1), and randomly assigned to the four parallel MST panels. The CPLEX is used for automated test assembly. The contribution of this study is two-fold. It will enable us a) to see the consequences of LID in MST, b) to choose best MST designs when local independency is violated.

Poster 29: Identifying Adaptive Algorithms for Increasing Comparative Judgement Efficiency: Review Study

San Verhavert, University of Antwerp; Vincent Donche, University of Antwerp; Sven De Maeyer, University of Antwerp; Liesje Coertjens, University of Antwerp

In 1993 Pollitt and Murray introduced the method of Comparative Judgement (CJ) in educational assessment. It is derived from Thurstone's Law of Comparative Judgement (1927) and is an alternative for marking (Pollitt, 2004). A group of assessors is asked to each individually make a set of judgements on which of two representations (e.g. essays) is a better representation of a competence under assessment. Using the Bradley-Terry-Luce model, a ranking from the lowest to the highest quality representation is estimated. Already from the early days it was observed that CJ needs a lot of of comparisons to reach an acceptable level of reliability (e.g. Bramley, Bell & Pollitt, 1998).

We will present a first step in tackling this inefficiency. In a systematic review we identify adaptive algorithms potentially increasing the efficiency of CJ and this from a broad range of research domains. In a first part, a taxonomy of the adaptiveness will be constructed. An exploratory review in the domain of CAT preliminarily revealed seven levels of distinction among which statistical paradigm (Frequentist v Bayesian), information measure based or not and weighting or balancing are the major distinctions. All distinctions and their meaning will be further elaborated on in the presentation. In a second part we will attempt to review the efficiency of these algorithms in their respective domain, keeping the CJ context in mind.

Besides informing CJ research this review might provide insights for (e.g.) Computer Adaptive Testing (CAT) research.

(CCC) CLASSIFICATION, CLUSTERING, AND LATENT CLASS ANALYSIS

Poster 30: Estimating Covariance in Longitudinal Mixture Modeling Using a Three-Step Approach

Ai Ye, University of Delaware; Luke Rinne, University of Delaware

An important issue in mixture modeling is the nature of relationships between latent class variables and auxiliary observed variables (Clark & Muthén, 2010). The standard way to estimate covariance effects is to simultaneously estimate the latent class structure and its regression on covariates in a single model called a "one-step" procedure. Such an approach is biased because covariance effects may influence latent class formation, potentially altering interpretations of class membership. This is especially problematic for longitudinal models in which class membership is interpreted uniformly over time (e.g., latent transition analysis; LTA). Three-step procedures have been proposed (Asparouhov & Muthén, 2014; Vermunt, 2010) in which the latent class structure is first estimated based only on the indicator variables; second, the standard error of most likely latent class membership is determined, and third, the final model including covariates is estimated using most likely latent class memberships as indicators, while controlling for the misclassification. In the current study, we conducted a Monto Carlo simulation in Mplus and R

(MplusAutomation) to investigate the performance of three-step method for LTA versus the one-step method under various conditions, including sample size, measurement quality (entropy value), and covariate effect size. We focus on differences in coverage and MSE when estimating different-sized covariate effects on latent transition probabilities. We then present an empirical study on the development of children's strategy use in fraction comparisons to illustrate how the three-step procedure can be used to model effects of covariates on transitions between latent class memberships over time.

Poster 31: Measurement Bias in Mixture Models: Results from an Experimental Study

Veronica T. Cole, The University of North Carolina at Chapel Hill; Daniel J. Bauer, The University of North Carolina at Chapel Hill

Mixture models, which aim to find meaningful subgroups of individuals based on patterns of indicators, have seen widespread use in recent years. For a number of phenomena (including depressive symptoms and alcohol use, among others) applications of mixture models have yielded inconclusive results such that the number and nature of classes uncovered by the models, as well as the predictors and outcomes of these classes, vary from one application to the next. This lack of clarity likely owes to a number of factors including differences in samples, study settings, and measures across applications. The current work formally explores the sensitivity of mixture model results to subtle differences in measurement using data from the Real Experiences and Lives in the University (REAL-U) study. In a unique laboratory analog study design, participants (N = 845) completed a battery of tests containing scales measuring a number of constructs, including alcohol use, substance use, and psychiatric symptoms. Subjects received both original items and "perturbed" versions of the same scale, in which item stems or response options were altered in order to simulate measurement differences across studies. Four experimental conditions, corresponding to different degrees of perturbation across individuals, were tested, permitting between- and within-person comparisons. Two mixture models, including a latent profile analysis of depression symptoms and a factor mixture model of alcohol expectancies, were fit under each of the four measurement conditions, and differences across conditions in the number and nature of classes were assessed.

(CDA) CATEGORICAL DATA ANALYSIS

Poster 32: Factor Analysis of Ordinal Items by Ridge GLS

Ge Jiang, University of Notre Dame; Ke-Hai Yuan, University of Notre Dame

Data in psychology are often collected using Likert-type scales, and it has been shown that factor analysis of Likert-type data are better performed on the polychoric correlation matrix rather than the product-moment covariance matrix, especially when the distributions of the observed variables are skewed. In theory, factor analysis of the polychoric correlation matrix is best performed using generalized least squares (GLS).

However, simulation studies showed that least squares (LS) or diagonally weighted least squares (DWLS) perform better than GLS, and thus LS or DWLS is routinely used in practice. In either LS or DWLS, the association among the polychoric correlation coefficients is totally ignored. To fill such a gap between statistical theory and empirical work, this article proposes new methods, called ridge GLS, for factor analysis of ordinal data. Results show that, for a wide range of sample sizes, ridge GLS methods yield uniformly more efficient/accurate estimates than existing methods (LS, DWLS, GLS). Rescaled and adjusted test statistics as well as sandwich-type standard errors following the ridge GLS methods also perform reasonably well.

Poster 33: 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.

(DIF) MEASUREMENT INVARIANCE AND DIF

Poster 34: Detection of Differential Item Functioning Based on Non-Linear Regression

Adéla Drabinová, Charles University in Prague; Patricia Martinkova, The Czech Academy of Sciences

Detection of Differential Item Functioning (DIF) has been considered one of the most important topics in measurement. Procedure based on Logistic Regression is one of the most popular tools in study field, however, it does not take into account possibility of guessing, which is expectable especially in multiple-choice tests. In this work, we present an extension of Logistic Regression procedure by including probability of guessing. This general method based on Non-Linear Regression (NLR) model is used for estimation of Item Response Function and for detection of uniform and non-uniform DIF in dichotomous items with presence of guessing. Proposed NLR technique for DIF detection is compared to Logistic Regression procedure and methods based on three parametric Item Response Theory (IRT) model (Lord’s and Raju’s statistics) in simulation study based on Graduate Management Admission Test. Non-Linear Regression method outperforms Logistic Regression procedure in power for case of uniform DIF detection and moreover by providing estimate of pseudo-guessing parameter. Proposed method also shows superiority in power at rejection rate lower than nominal value when compared to Lord’s and Raju’s methods based on three parametric IRT model. Our research suggests that the newly proposed non-IRT procedure which accounts for guessing is an attractive and user friendly approach to DIF detection.

Poster 35: How DIF Detection Accuracy Changes with Increased Dimensionality

Tyler Strachan, The University of North Carolina at Greensboro; Terry Ackerman, The University of North Carolina at Greensboro

Two commonly used approaches to detecting differential item functioning (DIF) are the Mantel Haenszel (MH) procedure and Sibtest. The goal of this research is to determine how sensitive the two procedures are to DIF as the non-valid dimensions increase. Specifically, in this study data will be simulated for a 50 item-test in which ten of the items display DIF. In all cases θ1 will be the valid dimension and the remaining dimensions will be the non-valid dimensions. That is, 40 items will have item vectors lying in a very narrow validity sector near the valid θ1 and the DIF items would be spread across the non-valid dimensions. Factors which will be varied include number of examinees in both the Reference and Focal groups (500 and 1000), number of dimensions (2, 3, 4 and 5), and correlations between dimensions (0, .5). In the two-dimensional case the DIF items will all have the same multidimensional discrimination (MDISC) and difficulty but will be measuring primarily the second (invalid) dimension θ2. In the three-dimensional case, five of the DIF items will measure primarily θ2, and five will measure primarily θ3. The pattern would continue for the 4- and 5-dimensional data. To create the DIF the mean values of the Ref and Foc groups will be centered at +.5 and -.5 respectively on each dimension with equal unit standard deviations. The correlations would be the same among the dimensions (either 0 or .5). Sibtest and MH analyses would be conducted on 100 replications for each of the 16 conditions.

Poster 36: DIF Detection for Passage-Based Items Using Dichotomous and Polytomous Scoring

Anita Rawls, The College Board; Xiuyuan Zhang, The College Board; Amy Hendrickson, The College Board

In reading and writing assessments, it is common to have several multiple-choice items relate to the same passage or stem. In tests, such as this, where there is a common stimuli, the assumption in many operational psychometric procedures is that there is no spurious correlation between the passage-based items introduced by the common stimuli. Consequently, the items are dichotomously scored and treated as if they are independent. However, if there is a spurious correlation between the passage-based items, polytomous scoring may be more appropriate. Ignoring the passage-induced correlation between passage-based items may cause an error in the item difficulty estimates, reliability estimates, and detection of differential item functioning (DIF; Bolt, 2002).

Using data from an administration of a large-scale, high-stakes reading and writing assessment for high school students, this paper compares the detection of gender DIF in passage-based items using the Mantel-Haenszel procedure (MHP; Dorans & Holland, 1993) and the generalized Mantel-Haenszel procedure (GMHP; Zwick, Donoghue & Grima, 1993). The MHP model uses dichotomous scoring while the GMHP uses polytomous scoring. The results are expected to provide guidance regarding the most appropriate approach for DIF detection for the passage-based assessment within a similar context.

Poster 37: Evaluation of Small Sample DIF Estimation Methods

Xiuyuan Zhang, The College Board; Anita Rawls, The College Board; Amy Hendrickson, The College Board; Timothy Moses, The College Board

Differential Item Functioning (DIF) analyses are essential in test development to ensure equality of test items to all demographic subgroups. When DIF is calculated using pretest data, one limitation that sometimes arises is that the pretest data may be collected with smaller sample sizes than desirable for DIF analyses. Our previous research showed that alternatives to standard DIF methods such as coarse or “thick” categorizations of the criterion score into quintiles and loglinear smoothing produced mixed results. Alternative approaches can usually improve the estimation accuracy but have somewhat different effects on multiple choice items compared to more difficult grid-in type items. The current project will build on the previous study. Pretest data from several forms of rights scored data, with items ranging from small-to-large amounts of DIF, will be evaluated. Sample sizes commonly encountered in pretest situations such as 50/200, 100/300, etc. for the focal and reference groups, respectively, will be considered using a real data simulation approach. Standard DIF analyses and alternatives such as thick matching and loglinear smoothing will be modified to address the limitations found in the previous study. An additional strategy based on logistic regression models that has been found to perform well with small sample data will also be considered. The intended outcomes for this project will be recommended systems, strategies and guidelines for the most accurate DIF strategies for pretest analyses.

Poster 38: A Comparison of DIF Detection Methods for Polytomously Scored Items

Ya Mo, National Institute of Statistical Sciences

This study was conducted to make an empirical comparison of differential item functioning (DIF) detection for polytomously scored items using Standardized Mean Difference (SMD) and Logistic Regression with residual analysis. The two methods were compared using eighth grade and twelfth grade students' performance on five prompts from the 2007 NAEP writing assessment. The two methods were employed to detect whether each prompt functioned differently on the performance of native English speakers and English as a second language (ESL) students. Results suggested that SMD and logistic regression with residual analysis tend to agree on the detection of DIF in polytomously scored items.

Poster 39: Factors Influencing the Mantel Method for Detection of Polytomous DIF

Elizabeth Patton, The University of North Carolina at Greensboro

The purpose of this present study is to provide a comprehensive table of the Type 1 error rate and power of the Mantel (1963) approach for detecting differential item functioning (DIF) in polytomous items with small sample sizes. A simulation study was employed in which polytomous data were generated using the Graded Response Model (GRM; Samejima, 1969). The Mantel approach was examined under four factors: (1) number of item response categories (3-,4-, and 5-), (2) degree of DIF (small and large), (3) in small sample conditions where the sample size of the reference and focal groups were equivalent and (4) in small sample conditions where the reference to focal group ratio was 4:1. In all cases, constant pervasive DIF was simulated, which exists when all of the steps underlying the polytomous response variable display DIF that is relatively equal in magnitude and direction (Penfield, Alvarez & Lee, 2009). For this study, the a-parameter was held constant while the b-parameter was adjusted by 0.3 and 0.6 for small and large DIF respectively. Sample size started at 100/100 (reference/focal) for the matched condition and 400/100 for the unmatched condition and increased incrementally until the size 2,400/2,400 and 2,400/600 were reached for the matched and unmatched conditions respectively. This was done in 10-stages for the matched condition and 6-stages for the unmatched condition.

Poster 40: Detection of Differential Item Functioning via a Bayesian Approach

Henghsiu Tsai, Institute of Statistical Science, Academia Sinica; Joyce Chang, Academia Sinica; Ya-Hui Su, National Chung Cheng University

Differential item functioning (DIF) occurs when individuals from different groups with the same level of ability have different probabilities of answering an item correctly. In this paper, we develop a Bayesian approach to detect DIF within the framework of item response theory models based on the posterior odds ratio. This procedure is compared to the Lagrange multiplier, the Mantel-Haenzel, and the logistic regression methods through simulations. An example of the data from the Department Required Test, which is a national test for applying colleges in Taiwan, will be presented to examine DIF on gender.

Poster 41: Simulation Combining Criteria and Comparing Degradation of Unidimensional DIF Indices

Jonathan D. Rollins III, The University of North Carolina at Greensboro; Amy Hendrickson, The College Board

Several non-IRT methods for detecting DIF are conditional on total scores, which may be confounded by the presence of both DIF and impact (i.e., group differences in performance). Methods which account for impact and/or multidimensionality have existed for quite some time (Cronbach, Ragosa, Floden & Price, 1977; Camilli, 1992). Still yet, traditional non-IRT methods provide a parsimonious solution (with fewer distributional assumptions and lower sample size requirements) and are widely used in operational settings. A simulation study is proposed that compares the degradation of traditional methods under fully-crossed conditions for sample size, number of true DIF items, uniform DIF amount, and level of impact. Dichotomous data are simulated from a two-dimensional, compensatory MIRT 3PL model to include a nuisance dimension to reflect the complexity of real data. Four DIF indices are compared: MH D-DIF, STD P-DIF, Breslow-Day trend, and logistic regression. Methods exploring combinations of flagging criteria are

studied as well. The primary purpose of this study is to determine the extent that combinations of criteria can be used in improving the detection of items which truly exhibit DIF. Results are evaluated by the number of DIF items correctly classified, Type I error, and Type II error. This study builds upon similar prior literature that explores this issue to further complement long-standing findings. Moreover, this study adds to the knowledge base by specifically suggesting heuristic approaches and combining flagging criteria which make connections between the simulation conditions and levels of error to further inform guidelines in using the indices.

Poster 42: Testing Measurement Invariance Across Gender Groups in Math Domain Abilities
Liuhan Cai, University of Nebraska-Lincoln

Measurement invariance is the essential psychometric property for scores to be comparable across groups. A lack of item parameter invariance, a special case of measurement invariance, will result in differential item functioning (DIF). This study used multi-group item factor analysis to examine the extent to which an item factor model measuring math abilities in domains of algebra, number, and geometry exhibited measurement and structural invariance between men and women. Data came from the Teacher Education and Development Study in Mathematics (TEDS-M). Dichotomously scored responses were obtained for 22 released items on the math content knowledge (MCK) assessment for the U.S. future secondary teachers. Results indicated that measurement invariance and structural invariance were established across gender groups in the domains of algebra and number. However, in domain geometry, only partial scalar invariance held due to the existence of a uniform DIF item. With the same ability level in geometry, the item was systematically more difficult for women than men. Furthermore, only partial residual variance invariance was achieved as another item from domain geometry contributed the largest source of misfit. The amount of variance in that item not accounted for by the ability in geometry was smaller for women than men. A diagnosis of the problematic items was conducted. This study demonstrated the need to establish construct validity for latent trait comparisons. Measurement invariance or at least partial measurement invariance must be established to ensure meaningful interpretation of group differences in math domain abilities.

Poster 43: Investigating Polytomous Item Invariance Using the LCDM
Oksana Naumenko, The University of North Carolina at Greensboro; Robert Henson, The University of North Carolina at Greensboro

Numerous differential item functioning (DIF) assessment methods have been developed in the frameworks of classical test theory (CTT) and item response theory (IRT) to address validity and fairness issues in testing. However, CTT and IRT DIF methods are not necessarily theoretically applicable with Diagnostic Classification Models (DCMs), which model the latent proficiency space as multidimensional and often dichotomous. The current study uses the Logistic Cognitive Diagnostic Model (LCDM; Henson, Templin, & Willse, 2009) as a framework for assessing DIF in polytomous items. The LCDM was selected because it is a general model that subsumes many of the known DCMs, such as the DINA model (Haertel, 1989; Junker & Sitjsma, 2001), and can be used with polytomous items (von Davier, 2014). Under this framework, item DIF is defined to occur when the probabilities of scoring at or below a particular score category differ for different groups of examinees with the same attribute profile.

The current study relies on simulation methods to compare the LCDM-DIF (Li & Wang, 2015) and the Mantel Haenszel (Mantel & Haenszel, 1959) for polytomous DIF in correctly identifying various types of DIF associated with polytomous items under the polytomous DINA model. Under the LCDM-DIF method, DIF is identified if the 95% confidence interval of the posterior distribution of the DIF parameter does not contain zero. Root mean squared error (RMSE), absolute bias, Type I error rate, and power will be examined for both DIF detection methods.

Poster 44: Geographically Weighted Item Response Theory for Differential Item Functioning Detection
Samantha Robinson, University of Arkansas

Mappings of spatially-varying Item Response Theory (IRT) parameters are proposed, allowing for visual investigation of potential Differential Item Functioning (DIF) based upon geographical location without need for pre-specified groupings. This proposed model merges Geographically Weighted Regression (GWR) and IRT methods, with the current emphasis being on a 1PL/Rasch model. This geographically weighted approach to IRT modeling and DIF detection provides a flexible framework, with various extensions discussed. Applications to simulated examination data, utilizing a kernel weighting scheme based on several fixed bandwidths as well as classically defined spatial neighborhood structures for both regular and irregular lattices, illustrates this method's benefit and practical value when comparisons are made to traditional DIF techniques. This approach, making use of three-dimensional surface mappings of estimated item difficulty parameters, serves to detect DIF across space without a priori groupings, thereby identifying regional disparities in item functionality that might be unobservable on a global level.

Poster 45: Multidimensional IRT Models be a Solution for Differential Item Functioning
Patricia Martinkova, University of Washington; Elizabeth A. Sanders, University of Washington; Yuan-Ling Liaw, University of Washington

This simulation study employs multidimensional framework to study differential item functioning (DIF) and to specify the psychometric dimensionality of the test items (e.g., math items in fact also measure language comprehension ability or communicating in writing, which could be deemed to be irrelevant or relevant to the math ability). We examine whether multidimensional item response theory (MIRT) models might be useful in controlling for DIF when estimating primary ability, or whether unidimensional IRT (UIRT) models with DIF items removed are sufficient for accurately estimating primary ability. The research question is addressed: item response data are simulated as 2PL two-dimensional noncompensatory items in which there is a low to modest discrimination parameter on the secondary dimension. How accurately do UIRT and MIRT models calibrate the primary ability estimate for focal and reference groups? Manipulated factors include proportions of items exhibiting DIF, relative proportion of focal and reference group sizes, and relative impact of the second dimension on item discrimination and difficulty. Each dataset is analyzed using five approaches: analyzing all items with a UIRT model; analyzing items with UIRT models after removing DIF items detected by MH and SIBTEST procedures respectively; and analyzing all items with two MIRT models. Result supports that MIRT models can be a practical tool to factor out DIF effects, particularly for tests with relatively high amounts of DIF. Although they aren't found to be significantly different, of the two types of MIRT approaches, the noncompensatory model averages slightly less bias than the compensatory approach.

(ECM) ESTIMATION & COMPUTATION METHODS

Poster 46: Model Violation Effects on Conditional Pseudo-Likelihood Polytomous Rasch Model Estimation
Saed Qunbar, The University of North Carolina at Greensboro; John Willse, The University of North Carolina at Greensboro

The proposed study will examine parameter recovery with polytomous Rasch models when using conditional pairwise pseudo-likelihood estimation with principal components (Andrich & Luo, 2003). This principal components Rasch model (PCRM) does not use the principal components analysis common in multivariate analysis. Instead, these principal components decompose polytomous items into a series of orthogonal polynomials (Guttman, 1950) associated with the spacing of thresholds. This technique has several advantages over Conditional Maximum Likelihood Estimation (CMLE). Willse, Rollins, and Qunbar (2015) showed that the technique recovers threshold estimates in small sample size conditions with less error than does CMLE. However, estimation of the PCRM using principal components was not tested in the presence of model violations. The proposed study will compare parameter recovery across 4 conditions; threshold, sample size, slope, and estimation. Data will be simulated using the Generalized Partial Credit Model (GPCM; Muraki, 1992). The threshold condition will be comprised of a 4 categories condition and a 7 categories condition. The sample size condition will contain conditions with sample sizes of 500 and 1,000. The slope condition will contain a condition with a common slope across all items (i.e., a standard Partial Credit Model; PCM; Masters, 1982) and a condition with unique slopes for each item (i.e., the GPCM). The estimation condition will compare the PCRM estimation of a PCM with a CMLE of the PCM, and Marginal Maximum Likelihood Estimation (MMLE) of the GPCM. The results will be evaluated by comparing estimated threshold and theta parameters with their true values.

(FAC) FACTOR ANALYSIS

Poster 47: Number of Response Alternatives: Implications for Factor Analysis Model Fit

Alexander G. Hall, University of South Carolina; Amanda Fairchild, University of South Carolina; Alberto Maydeu-Olivares, University of South Carolina

Previous methodological work has suggested that reducing the number of response alternatives on the same set of items may decrease the probability of rejecting an incorrect one-factor model using χ2 based fit indices (Green, Akey, Fleming, Hershberger, & Marquis, 1997; Maydeu-Olivares, Kramp, Garcia-Forero, Gallardo-Pujol, & Coffman, 2009). Relatedly, the power of the χ2 test statistic used in structural equation modeling decreases as the absolute value of excess kurtosis of the observed data increases. For discrete variables, the range that kurtosis can take depends on the number of categories, where excess kurtosis is more likely the fewer the number of categories. As a result, the fit of a factor analysis model to observed data is likely to improve with number of decreased response options on a scale, regardless of the underlying factor structure. To explore their phenomenon further, we extend previous work to formally evaluate this notion in a statistical simulation study where the impact of distributional nonnormality, model misspecification and/or improper model estimation are considered.

Results indicate that the effect of excess kurtosis (and hence number of categories) on the power of test statistics is compounded by model misfit: excess kurtosis has no effect when the model is correctly specified, but it increases as model misfit increases. Implications of these findings for substantive research is discussed.

Poster 48: Exploratory and Confirmatory Factor Analyses of the Pearman FlexIndex®

Gregory Gunn, Multi-Health Systems, Inc.; Joanna Solomon, Multi-Health Systems, Inc.; Vivian Wing-Sheung Chan, Multi-Health Systems, Inc.; Gill Sitarenios, Multi-Health Systems, Inc.

The Pearman FlexIndex® is a novel 37-item assessment that measures one's extent of psychological flexibility or agility. This study is the first to provide factor analytic results to support the validity of this assessment. Results are based on a demographically representative sample of 2,643 U.S. and Canadian employed adults who completed the assessment online. The sample was split for exploratory analyses (N = 1,322; 50.8% female with a mean age of 43.8 years [SD = 13.6 years]) and confirmatory analyses (N = 1,321; 50.8% female with a mean age of 44.0 years [SD = 13.8 years]). For exploration and item-reduction purposes, a series of factor analyses was conducted with the 61 proposed items using principal axis factoring with direct oblimin rotation and comparative data procedure (Ruscio & Roche, 2012). The pattern matrix, Eigenvalues, and scree plot results yielded a five-factor solution. Confirmatory factor analyses with the retained set of 37 items also supported a five-factor solution that accounted for 50.6% of the total variance in the model with satisfactory fit statistics. These five factors were best interpreted as Proactivity, Composure, Connectivity, Variety-Seeking and Rejuvenation. Furthermore, the model had minimal CFI value differences (CFI difference < .01) across gender, race and ethnic, and age groups, as evidence of measurement invariance (Cheung & Rensvold, 2002). Overall, the factor analytic findings support the validity of the Pearman FlexIndex®.

Poster 49: Constructs of the Korean Employability Skills Assessment and Work Ethic

HwaChoon Park, University of Georgia

The purpose of this study was to examine the constructs of work ethic measured by the Korean translation of the Employability Skills Assessment (KESA) testing the content validity and reliabilities of each construct of the KESA. The work ethic of Korean people was compared with two independent variables of generation and gender. Participants were Korean Baby Boomers (1955-1963) and Millennials (1982-2000) in South Korea. 473 respondents provided data during three weeks in summer in 2015 using the KESA. The Employability Skills Assessment (ESA: Hill, 1995) consists of 23 brief statements using a 7-point Likert scale to assess individuals' work ethic. The KESA is a Korean version of the ESA in order to provide Korean with a research-based instrument tested using statistical analyses and to enable Korean people to evaluate their own work ethic. Various factor analytic procedures including principal component analysis (PCA), and the common factor model including maximum likelihood (ML) with both orthogonal and oblique rotation methods, were performed to identify the construct of the KESA on data collected. The factors extracted were named. Reliabilities of each factor of the KESA were calculated using Cronbach's alpha. An ANOVA procedure was performed to compare the work ethic of Korean people based on generation and gender. Finally, effect size was calculated to interpret the significance of the results. The constructs derived provide understanding of east Asian culture related to work ethic.

Poster 50: Squared Canonical Correlations and Matrix Norms Under High-Dimensional Settings

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

In research on the closeness between factor analysis (FA) and principal component analysis (PCA), we typically compare FA loadings and PCA loadings using some measure of closeness or distance. Some studies have used the average squared canonical correlation between the two loading matrices as a measure of closeness. This measure has the advantages of being invariant with respect to sign and column changes, and most conveniently, it is not affected by orthogonal rotations. However, the drawback is that it is hard to intuitively perceive the amount of distance between the (elements of) two loading matrices. Therefore, use of the matrix norm(s) such as the Frobenium norm is sometimes preferred. However, then we encounter the complexities such as the sign changes and the column alignment of the corresponding factors/components as well as rotational indeterminacy. In the current study, the efforts are made to connect between the average squared canonical correlation and the matrix norms (e.g., Frobenium norm). In doing so, our focus is on the high-dimensional settings where the number of variables (p) should be taken into account as well as the sample size (N).

Poster 51: A Novel Method for the Modeling of Ordinal Count Items

Nathan Markiewitz, The University of North Carolina at Chapel Hill; Dan Bauer, The University of North Carolina at Chapel Hill

Many measures in psychology are obtained by asking individuals to report how often they engage in a behavior or set of behaviors. That exact rate or number may be difficult to recall, resulting in theoretically unexpected clusters of responses around convenient numbers. This phenomenon, known as heaping, is difficult to accommodate in subsequent data analyses. To sidestep this problem, many researchers use items that present subjects with intervals of counts (e.g. 0 times, 1-2 times, 3-5 times, etc.). Such "ordinal count" items are easy to include in surveys and avoid heaping by reducing cognitive demands. These advantages have led to their promotion by funding agencies and widespread implementation in large-scale longitudinal studies. A significant challenge that remains, however, is developing an optimal psychometric model for the unique features of ordinal count items.

Here we propose the ordinal count factor model (OCFM), which posits an underlying latent count process associated with each observed ordinal count item. The associations among these underlying latent counts are then explained by one or more common factors representing the measured construct(s). In addition to representing the data generating process more faithfully than a traditional ordered logit or probit factor model, OCFMs allow for inferences on the metric of the underlying rate of behavior. We demonstrate OCFM specifications for both bounded and unbounded count processes, with or without zero-inflation. Extensions to multiple populations will be discussed, including the advantages of OCFMs in integrative data analysis. Convenient options to estimate these models will also be presented.

Poster 52: Measurement Invariance of the Dark Triad Between Two Countries

Stephen Robertson, Clemson University; Cindy L. S. Pury, Clemson University; Jesus A. D. Datu, University of Hong Kong; Nino Jose Mateo, De La Salle University-Manila

We conducted a measurement invariance test of the 12-item Dark Triad Dirty Dozen measure (Jonason & Webster, 2010) between the United States (n = 203) and the Philippines (n = 291) using confirmatory factor analysis (CFA). The Dark Triad consists of three socially undesirable traits (i.e., Machiavellianism, narcissism, and psychopathy) that are characterized by dishonesty, a lack of empathy, and social manipulation. The traits routinely exhibit gender differences; therefore, a dummy coded gender variable was included in the CFA to control for gender at the item level. Configural invariance between the countries was supported. Next, the factor loadings and error covariances were examined for metric invariance. The Satorra-Bentler scaled $\chi 2$ difference test was nonsignificant for all constraints; therefore, metric invariance was confirmed between the two countries. Next, scalar invariance of the observed item means (intercepts) was examined. One Machiavellianism item, one psychopathy item, and two narcissism items were suggested to have significantly different intercepts across the two countries. The Machiavellianism and psychopathy items were higher in the Philippines, and the narcissism items were higher in the U.S. The possible response bias on these four items is not surprising because the agentic social style of the Dark Triad runs counter to the collectivist cultural expectation of selflessness (Jonason, Strosser, Kroll, Duineveld, & Baruffi, 2015). Overall, the Dirty Dozen measure exhibits metric invariance between the U.S. and the Philippines, but researchers should be aware of possible response bias on the indicated items.

(FCM) MODEL FIT, COMPARISON, AND DIAGNOSTICS

Poster 53: Simulation Study on Testing (Dis)ordinal Interactions Using Re-parameterized Regression

Ralph C.A. Rippe, Leiden University-Institute of Education and Child Studies

Individuals who vary in their responsivity to the environment can do so according to two major frameworks: a) diathesis stress (DTS), where some individuals are more susceptible to negative consequences of adverse experiences than others, and b) differential susceptibility (DS), where some individuals are more susceptible to both negative and positive consequences than others. Susceptibility markers include genetic, temperamental, and physiological factors.

Series of statistical tests have been proposed to evaluate the differential susceptibility model against the diathesis stress model. In essence, in moderation analysis these test an ordinal interaction against a disordinal one using re-parameterized regression, by centering the main predictor at its crossover point on the main predictor scale, instead of at the sample mean. In the ordinal case, the crossover point has to be equal to or greater than the highest observed point of the scale, while in the disordinal case the crossover point can lie anywhere within the observed range of predictor values. However, differences in sample size and (measurement) error are not accounted for, while their effects on both predictor and outcome side are evident in terms of attenuated effects and larger standard errors.

Through false positive rate and power calculations, we assess the effect of varying sample size and increasing error for predictor, moderator and outcome, on the ability to accurately test an ordinal against a disordinal interaction. Results show that, for all reasonable sample sizes, small but realistic amounts of error already obscure the detection of both effects and the competitive model evaluation.

Poster 54: Mapping Individuals onto a Multidimensional Emotion Space Using Self-Similarity Judgments

Jordan Sparks, Georgia Institute of Technology; James S. Roberts, Georgia Institute of Technology

A hybrid model approach to Carroll's hierarchy of preference models (1972) is presented to 1) provide a more parsimonious fit for preference judgments, 2) minimize the number of anti-ideal points that typically arise from External Multidimensional Unfolding (EMDU) models, and 3) guarantee that all model terms are statistically significant. The term "hybrid model" refers to situations in which the optimal regression model within Carroll's hierarchy has terms that are not all statistically significant, and consequently, such terms are eliminated. This elimination of terms from Carroll's original models leads to models in which alternative representations of preference may operate across stimulus dimensions. This methodology was grounded in the idea that there may be interpretable anti-ideal points in an EMDU solution, but they should account for a statistically significant amount of variation in the preference responses. This approach was applied to self-similarity judgments in the context of facial affect. Photos depicting facial emotions were scaled using multidimensional scaling of pairwise similarity judgments among photos, and then 1,564 subjects were located jointly in that same emotion space using self-similarity judgments. When the optimal model was selected for each subject, more than 95% of these models were hybrid models. Additionally, this new approach reduced the number of anti-ideal points by approximately 25% by allowing these points to become vectors in the group space. The results of this research illustrate that a hybrid approach to EMDU is an intuitive extension of Carroll's hierarchy that can represent preferences across stimulus dimensions in a less "all-or-none" fashion.

Poster 56: Bayesian Network Approach to Validate Item Mapping in Attribute Hierarchies

Seung Yeon Lee, University of California-Berkeley; Zachary Pardos, University of California-Berkeley

In cognitive diagnosis models, the specification of attributes that are measured by each item is fundamental to accurate assessment. In most cases, the association between attributes and items is defined by subjective judgment of domain experts, thus a risk of misspecification arises. This study proposes a Bayesian network approach to validate an attributes-to-items mapping, which is equivalent to the Q-matrix validation. We particularly consider the situation in which the attributes are hierarchically ordered (i.e., prerequisite relationship). It has been widely recognized that the attributes in mathematical and scientific concepts have hierarchical structures based on learning sequences in the curriculum. Existing methods of Q-matrix validation do not take into account such hierarchical nature of attributes. To the contrary, Bayesian networks are effective tools that can incorporate hierarchical relationships between attributes.

In this study, we provide a theoretical foundation of the proposed method. We further evaluate performance of our method by conducting two simulation studies and applying it to empirical data. Results show that our method can correct the item mapping for each item at a time under the assumption that the mappings of other items and associations between attributes are correctly specified.

(IRT) ITEM RESPONSE THEORY

Poster 57: Detecting Response Styles Using Discrepancy Statistics Within Multidimensional IRT

Daniel Adams, University of Wisconsin-Madison; Daniel Bolt, University of Wisconsin-Madison

Differences in how respondents use rating scales on self-report measures can undermine score validity. This study compares the performance of two IRT models and two discrepancy statistics evaluated using a Markov chain Monte Carlo method with posterior predictive checks. The discrepancy statistics are evaluated according to their ability to identify individuals who exhibit response style heterogeneity. We then compare the status of 463 respondents of the Wisconsin Inventory of Smoking Dependence Motives (WISDM-68) survey and of 394 respondents of the Enright Forgiveness Inventory (EFI) survey in terms of person fit.

Poster 58: Using Projective IRT Model in Test Equating

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

Stocking and Lord's (1983) test-characteristics-equating approach is a widely used equating method. The approach has shown advantages over other equating methods (Lee & Ban, 2009). However, for a test that is scaled using a single latent ability, local dependence and multidimensionality of the items could cause biases in estimating the students' latent ability (Li, Bolt, & Fu, 2005) as well as in equating, which depends on the accuracy of parameter estimation. The compensatory Multidimensional IRT (cMIRT) model for multidimensional data containing a dominant dimension and nuisance dimension(s) can be shown to be empirically indistinguishable to a projective unidimensional model (Ip, 2010). A projective IRT (pIRT) model which is derived from the cMIRT model can also correct for errors that are caused by the induced dependence of items (Ip & Chen, 2012). This enables the equating of tests that contain the same dominant but different nuisance dimensions.

The results of aforementioned studies suggest that the pIRT model will have better accuracy and lower standard error (SE) of the equated scores than the Testlet and the cMIRT model, and can be useful for equating tests with local dependency. In this study, the data from the Testlet model and the cMIRT model with a dominant dimension and a nuisance dimension will be simulated. The estimated parameters and latent abilities from the cMIRT model will be used to obtain those of the pIRT model. The recovery of equated latent abilities and their SE from Stocking and Lord's approach will be compared under various conditions.

Poster 59: Investigating Item Sensitivity to Response Style

Sien Deng, University of Wisconsin-Madison; Daniel Bolt, University of Wisconsin-Madison

Measurement of response styles (RS) in self-report rating scales have been investigated using the Multidimensional Nominal Response Model (MNRM). However, the assumption that all items are equally affected by RS may not be always true (i.e. RS loadings are equal across items). A recent MNRM (Falk & Cai, 2015) allows item slopes to vary across items for both substantive and RS dimensions, and can be used to test this assumption as well as investigate item characteristics that are associated with a reduced susceptibility to RS effects. In this paper, we first conduct a set of simulations to examine the recovery of item slopes of the RS dimension. Then we apply the model to real data to investigate the varying sensitivity of items in relation to different item characteristics. Different types of rating scale items from PISA 2012 student questionnaires are used for illustration. All analyses are conducted using Latent Gold 5.1 (Vermunt & Magidson, 2015).

Poster 60: WITHDRAWN

Poster 61: Non-Compensatory MIRT with Rotation

Xinchu Zhao, University of South Carolina; Brain Habing, University of South Carolina

Compensatory multidimensional item response theory (MIRT) models, allow for rotation of axes as in factor analysis. In those models, the correlation of the underlying abilities is arbitrary, and parameters from one correlation structure have a one-to-one link to any other. As such it is not necessary to specify or estimate the correlation to recover the model. This is not the case for the standard non-compensatory MIRT model due to its multiplicative structure.

The purpose of this study is to develop a non-compensatory MIRT model that allows for transformation between different correlation structures. This is accomplished through the use of additional discrimination parameters and allows for the estimation using an orthogonal ability distribution. The performance of the model was evaluated by simulation. In the simulation, the data was simulated for the standard two dimensional non-compensatory model with correlated abilities. It was estimated using the new model as if the abilities were uncorrelated. Parameters were estimated via Metropolis-Hasting algorithm within Gibbs sampler for both the standard non-compensatory and the new model. The parameter recovery was evaluated after transforming the estimates back onto the simulated scales where the abilities are correlated, using linking methods similar to those for the compensatory models. The parameters were well recovered.

Poster 62: A Review of PROC IRT in SAS/STAT

Hongwei Yang, National Board of Osteopathic Medical Examiners (NBOME); Hao Song, National Board of Osteopathic Medical Examiners; Kevin Kalinowski, National Board of Osteopathic Medical Examiners

This paper reviews the new Item Response Theory (IRT) modeling program PROC IRT in SAS/STAT. The procedure takes a factor analytic approach to IRT and treats an IRT model as a factor analysis model with one or more underlying common factors. The paper primarily examines unidimensional IRT models given heavy reliance on such models in applied research and operational work. To that end, multiple unidimensional IRT benchmark datasets including both dichotomous and ordinal polytomous responses are selected from the literature and estimated under PROC IRT. The benchmark applications are used to estimate several representative unidimensional IRT models (with/without parameter equality constraints) including Rasch, two-and three-parameter logistic, and two-parameter graded response models. The item and person parameter estimates are cross-validated with those from other IRT programs: WINSTEPS (binary responses only), R ltm package, SAS PROC NLMIXED and PROC MCMC with vague priors. The comparisons are made regarding consistency of parameter estimates, amount of computing time, availability of output information and licensing cost. The outlined process is repeated for several optimizers in PROC IRT for obtaining maximum likelihood estimates. Besides unidimensional IRT, the paper also briefly reviews the features of PROC IRT for multidimensional IRT, discusses the implications of the program for psychometrics research, for survey and test data analysis, and proposes areas of possible improvements. Finally, the paper provides the foundation for an extended review of PROC IRT through Monte Carlo simulations under various experimental conditions to investigate more aspects of the program: parameter estimation bias, etc.

Poster 63: Comparison of CTT and IRT in the Student Satisfaction Measurement

Yanhong Bian, Rutgers, the State University of New Jersey; Lu Wang, Rutgers, the State University of New Jersey; Chengbo Yin, Rutgers, the State University of New Jersey

Developed by Association of American Universities, Graduating Student Survey is widely used to evaluate students' satisfaction with their graduate programs and schools. Such evaluation is important for organization improvement and policy making. Traditionally, the survey is analyzed by Classical Test Theory (CTT). However, these analyses are not enough to provide knowledge of the overall quality of survey items.

This study aims to investigate whether Item Response Theory (IRT) can be appropriately applied to explore the item quality of Graduating Student Survey and at the same time provide different insights on the survey beyond the findings from CTT. Using the data of Graduating Student Survey obtained in 2013-2015 from the Rutgers University-Newark, an exploratory factor analysis was first conducted to explore the dimension of the survey. Then CTT and IRT were implemented to analyze the items within each factor. Both generalized partial credit model (Muraki, 1992) and the reduced two-parameter logistic model were employed to investigate the best fit of IRT models. Finally, logistic regression method was used to detect Differential Item Functioning between STEM and non-STEM students.

Finding suggests that although both CTT and IRT reveal similar information on items' contribution to the measurement of the latent trait, they provide very different insights on the quality and precision of the scale. Overall, IRT provides much richer information about survey items, and offers unique insights on how to improve the precision of the survey which will be highly valuable for research and practice in higher education.

Poster 64: Differencing Inattentive Respondents from Normal Respondents with/without Response Styles

Hui-Fang Chen, City University of Hong Kong; Kuan-Yu Jin, Hong Kong Institute of Education; Wen-Chung Wang, Hong Kong Institute of Education

Self-report assessments assume that all participants are highly motivated to accurately respond research questions. However, previous studies have identified 5% to 50% of participants who lacked of motivation to answer question. The present study proposed a mixture item response model to differentiate unmotivated respondents, who endorse the given alternatives randomly or consistently choose the middle-point response category, from normal respondents, who indeed pay attention to survey questions. Also, the new model divides normal respondents into two groups with distinct two latent traits: the intended-to-be-measured proficiency, and the tendency to endorsing middle (e.g., response categories of 2, 3, and 4 on a 5-point scale) or extreme response categories. A series of simulations were conducted to evaluate the parameter recovery of the new model and the generalized partial credit model (GPCM), which does not take into concern inattentive respondents and participants with specific response styles (RS). Different proportions of inattentive versus normal respondents were manipulated. It showed that, when the generated data mixed normal (with and without RS) and inattentive respondents, the proposed model had a high precision rate in classification of inattentive and normal respondents and recovered item and person parameters fairly well, whereas the GPCM yielded underestimated slope and difficulty parameters and shrunken thresholds. Moreover, when there were all normal respondents with RS, the proposed model still performed very well, whereas biased parameter estimates were derived from the GPCM. An empirical example was provided to demonstrate the applicability of the new model.

Poster 65: Technique to Reduce Data Demands of an Unfolding MIRT Model
Elizabeth Williams, Georgia Institute of Technology; David R. King, Georgia Institute of Technology; James S. Roberts, Georgia Institute of Technology

The current study explores a technique to reduce the data demands incurred when using the Multidimensional Generalized Graded Unfolding Model (MGGUM; Roberts & Shim, 2010). The MGGUM is a highly parameterized unfolding item response theory (IRT) model that generally requires large samples relative to most psychological research methods. This research implements a two-step technique in which multidimensional scaling (MDS) of pairwise similarity judgments is used to estimate stimulus locations, and then these locations are held fixed when single stimulus responses are analyzed with the MGGUM. All other MGGUM person and item parameters are estimated in this step other than stimulus locations. The effect of using fixed item locations derived from MDS is evaluated using a simulation study. Two dimensional MGGUM data are generated using three levels of sample size (N=500, 1000, 1500) crossed with two levels of stimuli (15 and 30 test items). The MGGUM will be estimated using the Metropolis-Hastings Robbins-Monro (Cai, 2010) algorithm with and without fixing the item location parameters to their MDS estimates. Recovery of true parameters will be examined for both solutions. The number of subjects needed for an MDS solution is substantially less than that needed for IRT calibration. Therefore, we hope to reduce the demands of the MGGUM by using the MDS solution to constrain item locations. If successful, then the applicability of the MGGUM will increase in situations where the number of subjects is less than traditionally needed for IRT.

Poster 66: WITHDRAWN

Poster 67: Scoring Algorithm for Students' Constructed Responses
Tomoya Obuko, The University of North Carolina at Greensboro

In this research we present a scoring algorithm for essay type items using latent class modeling and natural language processing. Recently, student-produced response question, such as essay-type questions, is attracting test administrators' attention with increasing number of Computer-based testing. Generally, automated scoring algorithm scores essays / short-answer questions based on some statistical indices with an assumption of uni-dimensionality in a factor. On the other hand, our presenting model assumes multi-dimensionality in the scoring. Some scores are predicted for each essay, because the presenting model allows us to have several patterns of scoring criteria even in one-factor model. The patterns are described as latent classes. One of the classes is selected as the final score for the essay based on the other criteria or external variables; therefore, only one score is presented to the answer eventually. In this presentation, we show some results on the presenting scoring algorithm that is used for some university entrance examinations, and evaluate reliability of the algorithm. Then we discuss possibility and feasibility of the scoring algorithm, especially in large-scale testing.

Poster 68: A GLMM Reformulation of the 2PL Model with MCMCglmm
Menglin Xu, The Ohio State University

The purpose of the study is to show how a reformulation of the two-parameter logistic model (2PLM) is achieved within the framework of GLMM, including the conversion equations to obtain the common 2PLM parameterization, an application with real data, and a simulation study to investigate the goodness of recovery with the R package MCMCglmm. Our real data study showed that for the item intercept, the correlation between the converted MCMCglmm estimates and flexMIRT results is .999, and for the item slope, the correlation is .909, indicating good parameter recovery. It is followed by a small simulation study, which is a 2x2x2 design with 20 replications per cell: 20 or 40 items; sample size (N) = 200 or 1000; slopes are distributed based on a lognormal with mean = 0.5 and sd= 0.2 or 0.4; item intercepts and person abilities are all normally distributed. The conversion equations, results of real data applications and the small simulation study will be presented in the conference. Its contribution to methodology innovation will be discussed.

Poster 69: Item Parameter Hyperpriors in Item Response Theory (IRT)

Jessica N. Jacovidis, James Madison University; Allison J. Ames, James Madison University

In Bayesian item response theory (IRT), parameters of the prior distribution- hyperparameters- can have priors themselves, termed hyperpriors. Hyperpriors are used to reduce prior informativeness, perform noninformative analysis, and allow observed data to influence the hyperparameters (Stone & Zhu, 2015). Additionally, hyperpriors can provide robust results by borrowing, or pooling, information across observational units (e.g., persons or items) and can result in more stable Bayesian estimates (Johnson, Sinharay, & Bradlow, 2007; Fox, 2010).

Hyperpriors give rise to shrinkage estimates of IRT item parameters, where the amount of shrinkage toward the prior mean is inferred from the data (Fox, 2010). However, there is little to guide practitioners on factors affecting the amount of shrinkage in IRT item parameters, and when the use of hyperpriors can improve parameter estimation. Sheng (2013) examined the effects of hyperpriors on discrimination and difficulty parameters; however, use of hyperpriors on the guessing parameter has not been examined. To address this gap, this study will focus on evaluating the use of hyperpriors on the guessing parameter.

Data will be simulated for various test lengths (10, 20, and 100 items), sample sizes (100, 300, 1000, and 5000 people), and test purposes (tests measuring a broad range of ability and tests used for cut score decisions). Item parameter recovery bias and posterior coverage rates for models with hyperpriors will be compared to models without hyperpriors (using both noninformative and informative priors). Recommendations for practitioners will also be offered.

Poster 70: Fitting Graded Response Models to Data with Nonnormal Latent Traits

Tzu Chun Kuo, Southern Illinois University-Carbondale; Yanyan Sheng, Southern Illinois University-Carbondale

Fitting item response theory (IRT) models often relies on the assumption of a normal distribution for the person latent trait(s). Violating this assumption may bias the estimates of IRT item and latent trait parameters (e.g., Sass, Schmitt, & Walker, 2008; Reise & Revicki, 2014) especially when sample sizes are not large. In practice, the actual distribution for person parameters may not always be normal, and hence it is important to understand how IRT models perform under such situations. This study focuses on the performance of the multi-unidimensional graded response model using a Hasting-within-Gibbs procedure (Kuo & Sheng, 2015). Two dimensions are considered in this study with the intertrait correlation being 0.2, 0.5, or 0.8. Model parameters are estimated for datasets with 500 or 1000 persons' three-scale Likert-scale responses to 20 or 40 items, where four distributions of latent traits are considered. These distributions have the following specified skew and kurtosis (Headrick, 2010): (1) skew = 0, kurtosis = 0 (normal), (2) skew = 0, kurtosis = 25 (symmetric and heavy tailed), (3) skew = 2, kurtosis = 7 (slightly skewed and heavy tailed), and (4) skew = 3, kurtosis = 21 (more skewed and heavy tailed). The results of this study provide a general guideline for estimating the multi-unidimensional graded response model under the investigated conditions where the latent traits may not assume a normal distribution.

Poster 71: Application of Testlet Models in Multidimensional Rasch Model

Dan Wei, Beijing Normal University; Danhui Zhang, Beijing Normal University; Hongyun Liu, Beijing Normal University; Peida Zhan, Beijing Normal University

Different types of testlet models have been widely used in education assessments in that the items constructed based on a common stimulus are subject to the assumption of IRT local independence. Most of the previous researches have investigate the nuisance effects caused by testlet traits under the condition that only one dimension trait θ was modeled. This study aimed to explore the effects of testlets in multi-dimentional Rasch models with the marginalized maximum likelihood estimation (MMLE) in Conquest. Two simulation studies and one real data analysis were included. Study I explore the estimation accuracy of models parameters in different structure and different degrees of complexity. In study II, three factors were considered: the number of items in each ability dimension, the number of items in each testlet, and the correlation among different abilities. Two multidimensional Rasch model with and without testlet effects were compared in terms of estimation bias, RMSEA, percent coverage of 95% confidence interval, as well as the reliability of ability θ estimation. It was discovered that as the number of items in each ability-dimension and the number of items in each testlet decreased, the item parameter estimation biases increased. Regarding with the real data, three dimensional abilities in math were modeled combined with two testlet effects. Overall, it also showed that there were great differences between the simple MIRT model and the one that explicitly took the testlet structure into account, which suggested that the non-independence due to testlets was a larger problem on this test.

Poster 72: Quasi Monte Carlo Method in Item Response Theory Estimation

Jichao Zhong, Beijing Normal University; Tong Ran, Beijing Normal University; Danhui Zhang, Beijing Normal University

Item Response Theory (IRT) is widely adopted for academic achievement assessment. However, under certain circumstances, like huge sample size and complex model structure, slow convergence or no convergence often occurs, which called for improvements in computation. This study discussed the application of Quasi Monte Carlo (QMC) method and its node setting in Multidimensional Model of Item Response Theory, with Maximum Likelihood Estimation (MLE) as the estimation method. Software used in the study was the TAM package in R; and Data was from the mathematics tests of National Assessment of Education Quality (NAEQ) in China. Evaluation and comparison were made through the correlation of key parameters (item difficulty and student ability), processing speed and model fit (AIC, BIC, deviance). Results indicated significant difference between models with and without QMC, and among models with various node settings. Key parameters were highly correlated with or without QMC, but the QMC model outperformed in Standard Error of Estimation and processing speed, as well as major model fit indexes. The application of QMC can serve as an augmentation to the accuracy and efficiency for NAEQ assessments.

Poster 73: An Extention of Rudner-Based Consistency and Accuracy Indices for MIRT

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

Although multidimensional item response theory (MIRT) has enjoyed tremendous growth, solutions to some problems remain unavailable. One case in point is the estimate of classification accuracy and consistency indices. Yao (2013) and LaFond (2014) focused on accuracy and consistency indices under MIRT based total sum scores only. It is problematic because of two reasons⍰for one thing, classifications made with the latent ability estimates shall be equally or more accurate than classifications made with total sum score (Lathrop & Cheng, 2013), at least for graded response model; and for another, it may be difficult to estimate accuracy and consistency indices in each content areas when some items may measure more than two domains (complex structure). Guo-based index has been extended to MIRT under complex decision rule (Wang et al., 2015, Paper posted at the 80th Meeting of the Psychometric Society). The purpose of this proposal is to extend Rudner-based index for MIRT, and to compare it with Guo-based index. Rudner-based index assumes estimation error is multidimensional normally distributed around each examinee's estimate of θ, and a simple Monte Carlo method can then be easily used to estimate accuracy and consistency indices. The simulation study was conducted to investigate whether the new index can work well under various conditions. Finally, the applications and directions based on the current research are suggested.

Poster 74: Evaluating the Triarchic Psychopathy Measure: An Item Response Theory Approach

Yiyun Shou, The Australian National University; Martin Sellbom, University of Otago; Jing Xu, Affiliated Nanjing Brain Hospital of Nanjing Medical University

Recent efforts have been made to introduce and operationalize the construct of psychopathy in East Asian countries. Most previous studies validated and evaluated the translations of measures of psychopathy by relying on the classical test theory approach. The analyses provided limited item level information especially in the cross-cultural context. In this study, we presented an item response theory analysis of the Chinese version of the Triarchic Psychopathy Measure (TriPM). Graded response models were used to analyze the three scales of the TriPM. We evaluated the item information and the range of the latent trait assessed by the TriPM. We identified the items that were not effective in the Chinese context. The results also revealed that some cultural factors, such as response style and perception of the wording valence, could have impact on dimensionality and local independence of the scales. We will discuss the implications and future directions for applying IRT models in evaluating measures of psychopathy.

Poster 75: Evaluation of IRT Scores for IDA Using a Within-Subjects Design

Michael L. Giordano, The University of North Carolina at Chapel Hill; Daniel J. Bauer, The University of North Carolina at Chapel Hill; Andrea M. Hussong, The University of North Carolina at Chapel Hill; Patrick J. Curran, The University of North Carolina at Chapel Hill

Integrative Data Analysis (IDA) allows for the fitting of statistical models to aggregated data drawn from two or more independent samples. A central challenge in IDA is developing harmonized measures based on multi-item scales differing in item stem or response across studies. Although IDA is becoming used with increasing frequency, assumptions underlying the process of psychometric harmonization have yet to be carefully studied. While analytic simulations can provide insights into some issues, neither approach evaluates how actual people respond to actual items. To address this, we collected data on 854 participants using a within-subjects experimental design; participants responded to a variety of scales that differed in item stems and response options in highly structured ways. This allowed us to approximate the counter-factual of how an individual would have responded had they been presented with two different versions of the same scale. For the current paper we focused on responses drawn from two variations of an alcohol expectancies measure. We fit a multivariate explanatory IRT model to obtain factor scores on both versions of the measure for a subset of participants. Test-retest correlations for the two versions ranged from 0.78-0.85 for simple mean scores and 0.73-0.85 for IRT factor scores. More importantly, correlations between the two versions across time yielded correlations of 0.69–0.74 indicating a high degree of overlap in responses between the two versions of the same scale. These results support the feasibility of current scoring methods in IDA and suggest novel directions for future work on improving score estimation.

Poster 76: Bootstrap Equating Errors for the NEAT Design Using Rasch Methods
Mingying Zheng, University of Nebraska-Lincoln

In most large-scale testing programs, more than one form of a test is often administered at different times and at different locations. The nonequivalent groups with anchor test (NEAT) equating design is traditionally based on using a single anchor to adjust for differences in test difficulty which is critical to equating test forms in most large-scale testing programs. Practitioners might face a trade-off between maximizing the anchor length for statistical purposes and minimizing it for other considerations such as test security and item datedness. When tests differ somewhat in content and length, methods based on the item response theory (IRT) model lead to greater stability of equating results. The current study extended the research of Tsai, Hanson, Kolen, and Forsyth (2001) by comparing standard errors, bias, and root mean square errors using four Rasch IRT equating methods for the nonequivalent groups with anchor test design: (a) whether Rasch parameters are estimated separately or concurrently, (b) whether a scale transformation is calculated to place parameters for the two forms on a common scale (e.g., Stocking & Lord, 1983, method), (c) whether the true-score or observed-score equating method is used. The sizes of the equating anchor were employed in all four different Rasch equating methods to investigate how different anchor sizes may impact the test accuracy of the tests by conducting a simulation study.

Poster 77: Ordered Partition Model for Confidence Marking Modeling
Oliver Prosperi, Université de Fribourg

Confidence marking is increasingly used in multiple choice testing situations, but when a Rasch measurement model is applied to the data, only the binary data is used, discarding all the information given by the confidence marking. This study shows how Wilson's Ordered Partition Model (OPM), a Rasch family model, can be used to model the confidence information and provide a powerful diagnostic tool to assess item difficulty, overconfidence or misuse of confidence levels but also the fact that a question is particularly tricky or creates a lot of doubt.The result is a model, which is in strict relation to the binary Rasch model, since the Rasch ICC's are "split" into a set of curves each representing a confidence level. The new model provides a set of item parameters that map the probability of being in each confidence level in relation to the test-taker's latent ability trait.

Poster 78: New Method for Modeling Directional Local Item Dependence

Kaiwen Man, The University of Maryland-College Park; Hong Jiao, The University of Maryland-College Park; Meng Qiu, University of Maryland-College Park

Local item dependence (LID) is likely to be present when a test is constructed of items that are clustered around a testlet. A testlet (Wainer & Kiely, 1987) is defined as a common stimulus where multiple items are constructed based on like in a passage-based reading comprehension test. In testlets, local item dependence is most often non-directional. Another scenario is the multipart items where the answer to the second part of an item depends on the answer to the first part of the item. This directional local item dependence is less studied in literature.

This study proposes a new approach to accounting for directional local dependence in multipart items. The new method models directional local item dependence between two items by estimating the correlation of difficulty parameters of two items that are dependent by assuming the bivariate normal relationship among the difficulty parameters. Markov Chain Monte Carlo method is used for model parameter estimation. Model parameter recovery is evaluated in a simulation study in terms of item and ability parameter estimation.

(LDA) LONGITUDINAL DATA ANALYSIS

Poster 79: An Application of State-Space Time Series Analysis to Psychophysiological Data

Daniel M. Smith, The University of Rhode Island; Theodore A. Walls, The University of Rhode Island; Mohammadreza Abtahi, The University of Rhode Island; Kunal Mankodiya, The University of Rhode Island

State-space modeling is a framework for time series analysis in which it is possible to execute many types of models for intensive longitudinal data such as ARMA, time-varying regression, multilevel modeling, and dynamic factor analysis. Its flexibility, computational efficiency, and handling of multiple observed and latent variables and missing or unevenly spaced observations make it a convenient approach for psychophysiology settings, in which small-N designs involving thousands of observations per participant are common. Recent scholarship by Schuurman et al. (2015) highlights the distinction between measurement error, which affects the current observation only, and dynamic error or innovations, which carry over to subsequent measurement occasions. These authors demonstrate the bias in parameter estimates that results from AR(1) models, which disregard measurement error, and posit two alternative models that account for measurement error. These models include AR(1) plus white noise and ARMA(1,1). In this study, we consider the strengths and weaknesses of these models against competing formulations. We demonstrate the application of these models using state-space time series analysis to model heart rate data collected during a cognitive task in patients with a history of suicidality.

Poster 80: Behavioral Trajectories of Young Children from Preschool to Grade One

Jin Liu, University of South Carolina-Columbia; Christine DiStefano, University of South Carolina

The purpose of this study is to investigate the behavioral trajectories of young children from preschool to First Grade. We tracked 233 children's behavioral problems from preschool to First Grade from fall, 2011 to spring, 2014 (6 time points). Two studies are included.

Study 1: In each time point, children's behavior was rated using the behavioral and emotional screening systems (BESS), and a risk status (Normal or At Risk) was identified. PROC GLIMMIX in SAS ® 9.4 is used to analyze the probability of being identified as at risk across time points and how this relates to the demographic characteristics (gender, lunch status, and English Language status).

Study 2: In addition to the overall risk status, children's subscale problems scores (externalizing problems, internalizing problems, and adaptive skills) were obtained at each time point. Mplus 7.2 is used to analyze the longitudinal invariance of the subscales in the CFA framework. A series of models is tested to build the measurement invariance. Model fit (chi-square values, CFI, TLI, and RMSEA) and difference testing are investigated. If the longitudinal invariance is built, the follow-up study will focus on investigating the behavioral changes across time for three subscales and how this relate to the demographic information.

Both studies will offer researchers a better understanding of young children's behavioral changes.

Poster 81: Challenges and Strategies for Estimating Second-Order Latent Curve Models

Stephanie Lane, The University of North Carolina at Chapel Hill; Kenneth Bollen, The University of North Carolina at Chapel Hill; Samuel McLean, The University of North Carolina at Chapel Hill

In the last two decades, latent curve modeling (LCM) has been used extensively across various domains within the social and behavioral sciences to investigate stability and change over time. Most frequently, researchers have employed latent curve models in which a single observed variable serves as the repeated measure at each wave of measurement. However, when multiple indicators of a given construct are present at each wave, we may instead use these manifest variables as indicators of a first-order latent variable. In turn, the first-order latent variables constructed at each time point may then serve as indicators of higher-order growth factors. This parameterization is known as the second-order LCM, or the curve-of-factors model (CUFF; McArdle, 1988). Many potential benefits are associated with this approach, such as the ability to assess measurement invariance and the ability to disentangle measurement error. Despite these advantages, the second-order LCM has been implemented relatively infrequently since its introduction more than two decades ago.

Here, we examine a variety of challenges researchers may face in fitting a second-order LCM. For instance, improper solutions and nonconvergence are more frequent when many indicators exist at each measurement occasion. In addition, the complexity of the model specification increases with the introduction of more parameters. We examine strategies for building a second-order latent curve model in the presence of a large number of manifest variables. Finally, we demonstrate these strategies on empirical data from a prospective cohort study examining outcomes associated with the recovery process following a motor vehicle collision.

Poster 82: Accounting for the Initial-Growth Relationship in Group Difference on Growth

Xiaoyan Xia, University of Pittsburgh; Feifei Ye, University of Pittsburgh

Educational analysts studying changes in achievement and other educational outcomes among different groups of students are presented with many challenges in developing a statistical model to compare growth among groups. With only two time points of data, methodologists have argued it may be difficult to make robust inferences concerning differences in achievement growth. In the presence of an association between initial status and growth (e.g., due to measurement error), and when the groups differ on initial status, the choice of modeling specification can affect the estimation of group differences in growth (Allison, 1990). When three or more time points of data are available, advances in statistical modeling have made it possible to overcome these challenges. This study examined the possible advantages of a three-time point design over a two-time point design in examining the group difference in growth. We illustrated the consequences of intrinsic associations between initial status and growth and measurement error on the estimation of group difference in growth with different statistical models, including change score model, regressor variable model, and latent variable regression (Raudenbush & Bryk, 2002). The simulation factors include measurement error size (reliability=.6, .7, .8, .9, .95, and 1.0), the initial-growth relationship (standardized slope=0, .2, .5, .8), and initial group mean difference (Cohen's d =0, .05, .1, .2). Results from this study will help guide applied researchers in terms of whether to include additional time points and which model to select, given the challenges of various measurement error, initial group difference and the initial-growth relationship.

(MDS) MULTIDIMENSIONAL SCALING

Poster 83: Individual and Group-Level Multidimensional Scaling of Physical Attraction Data

Matthew E. Barrett, The Georgia Institute of Technology; Eliana J. Dubin, The Georgia Institute of Technology; James S. Roberts, The Georgia Institute of Technology

This paper investigates several different properties of scaling physical attraction data at both the group and the individual level. Digital movies of the same female model were created by a graphic designer that varied characteristics of the model with respect to waist, hip, and bust sizes. These characteristics were nested within overall weight to create 81 total stimuli. Group-level multidimensional scaling was accomplished by assigning randomized blocks of pairwise stimuli; every 20 blocks constituted full pairwise data for one "virtual subject". All participants completed pairwise similarity judgments, attractiveness judgments, and demographic questionnaires. A 4 dimensional solution was obtained at the group level, confirming the overall salience of the manipulated characteristics of the female models. Separate pairwise similarity matrices were obtained for each virtual subject and an INDSCAL analysis was conducted on the data. Variation in the salience of the individual dimensions was detected and related to the demographic characteristics via multivariate regression. The variance of the salience of each of the physical dimensions was most closely associated with the self-reporting of the presence of a female guardian in the house while growing up. The efficacy of conducting group-level and individual-level multidimensional scaling of data that is aggregated from multiple sources is discussed.

Poster 84: WITHDRAWN

(MIS) MISSING DATA

Poster 85: Stratified Doubly Robust Estimator when Data are MAR or MCAR

R. Wes Crues, University of Illinois at Urbana-Champaign; Carolyn J. Anderson, University of Illinois at Urbana-Champaign

This study explores a stratified doubly robust (sDR) inspired estimator when data are assumed to be missing completely at random (MCAR) or missing at random (MAR). Hattori & Henmi (2014) proposed a stratified doubly robust estimator for the average treatment effect, where the estimator is stratified based on the ratio of two possible propensity score models. We extend their approach to instances when data are missing. The sDR approach allows two opportunities to propose logistic regression models to estimate the probability an observation is missing, along with an outcome regression to estimate a plausible value for the missing value, by the generalized linear model, and combines these to estimate the location parameter of a statistical distribution. Here, we stratify on the ratio of the estimated probabilities of missingness. We simulated data from a multivariate normal distribution and a multivariate skew normal distribution (Azzalini & Valle, 1996). From the simulated data, we created missing data at varying proportions in one column of our data, to which we used the sDR method to estimate the location parameters. We compared the results of the sDR method to multiple imputation in terms of bias and efficiency, and found an interesting phenomenon; namely, the proposed estimator's performance with respect to bias and efficiency yields more accurate and efficient estimates of the location parameter as the proportion of missingness increases. We posit this is a result of our method to compute the probability of missingness.

(MLM) MULTILEVEL/HIERARCHICAL/MIXED MODELS

Poster 86: Robust Standard Errors for Non-Normal Residuals in Multilevel Modeling

Lida Lin, University of Pittsburgh; Feifei Ye, University of Pittsburgh

Robust standard errors are commonly used in multilevel modeling in the presence of non-normal errors and/or heteroscedastic errors. Research has shown that robust standard errors work for non-normally distributed level-2 errors only when sample size is large enough (level-2 sample size greater than 100). However, not all empirical studies followed this rule. This study extends previous studies (Maas & Hox, 2004; Hox & Maas, 2011; Meuleman & Billiet, 2009) on the performance of robust standard errors by examining the impact from both level-1 and level-2 non-normal errors, with a focus on small level-1 and level-2 sample sizes. Simulations varied 1) non-normal residual distributions (chi-square, uniform, Laplace) in both level-1 and level-2; 2) intraclass correlation (0.1, 0.2, 0.3); and 3) sample sizes (level-1: 5, 30, 50; level-2: 30, 50, 100). Results show that non-normal residuals at level-1 and level-2 have little impact on the parameter estimates of fixed and random effects. For the fixed parameters, non-normal level 1 and level 2 residuals resulted with unbiased standard errors for the effects involving the level-2 predictor, but overestimated standard errors for the intercept and the level-1 predictor, even though robust standard errors performed relatively better with a smaller bias. When the non-normality is at only level 1, the robust standard errors tend to be overcorrected and more biased, especially for smaller number of groups. Larger sample sizes, in particular at level-2, help reduce the standard error bias for both ML and robust estimates.

Poster 87: Multilevel Model Selection and Predictor Rankings Using Criticality Analysis

Razia Azen, University of Wisconsin-Milwaukee; Luciana Cancado, University of Wisconsin-Milwaukee

In this study we evaluate the use of criticality analysis for the selection of a model and ranking of predictor importance in multilevel models. For a given sample, criticality analysis (Azen, Budescu, & Reiser, 2001) fits all subset models possible from the predictors of interest and rank-orders the models using an appropriate measure of fit. The analysis then employs the bootstrapping procedure to obtain the distribution of “Best Fitting Models” over many samples, thus taking sampling error into consideration in identifying the model that is likely to fit best. Predictors that appear more frequently in the best models are considered critical to identifying this model and are ranked more highly. The criticality of a predictor is thus defined as the proportion of best-fitting models that included that predictor. Through simulation, we varied the correlations within (level-1) and between (level-2) groups, sample sizes, intra-class correlations, and types of multilevel models to determine whether criticality analysis can identify the true (or best) model as well as provide an accurate rank-ordering of predictors. Result indicate that criticality analysis is suitable for both purposes under most conditions and with several measures of fit. The rate of correct model identification and accurate rank-ordering of predictors was found to be high when the number of observations per group was greater than 10 or when there were more than 10 groups. The results inform the development of recommendations regarding the baseline levels of predictor criticality and the general use of criticality analysis for multilevel modeling.

(PCC) PRINCIPAL COMPONENTS & CORRESPONDENCE ANALYSIS & RELATED METHODS

Poster 88: A Review of Extensions of the Optimal Coordinate Approach

Theodore Walls, The University of Rhode Island; Valerie Ryan, The University of Rhode Island; Daniel Smith, The University of Rhode Island; Gilles Raîche, Université du Québec à Montréal

The optimal coordinates approach to determining the number of components to retain was developed in the mid-2000s by Raîche and several collaborators. Following in depth coverage of the problem in the journal Methodology (Raîche et al., 2013), the approach has seen fairly diverse application in a range of sciences, and a limited amount of discourse has emerged around relative strengths and its further development. For example, applications have emerged in areas as diverse as neuroscience, power generation monitoring, genetics, and in several scale development projects. Some consideration of model implementation has emerged as applications have variously applied competing methods in different situations and experimented with different decision-making criteria, and some researchers have looked into model refinements. This paper takes a closer look at selected applications, considers themes that have emerged around further evolution of the method, and considers options for further refinements.

(PRO) PATIENT-RELATED OUTCOMES

Poster 89: Linking Between PROMIS and Legacy Instruments: GRM Versus Rasch PCM

Man Hung, University of Utah

In the past few years, many studies have been conducted to examine the performance of the Patient-Reported Outcomes Information Systems (PROMIS) instruments, yet there is little research establishing the comparability of the PROMIS instruments to legacy instruments measuring similar constructs. If a variety of instruments are used in research, there can be a considerable amount of challenges for clinicians and researchers to understand and compare results from different studies. Recognizing such gap, the aim of this study was to conduct a cross sectional study to establish comparability of the PROMIS instruments and legacy instruments in measuring functional status. Both the Graded Response Model and the Rasch Partial Credit Model were separately applied to perform the linking procedure. Two cross walk tables were produced matching each summed score point on a legacy instrument to a PROMIS T-score and their metrics were compared to identify discrepancy from results generated by the Graded Response Model and the Rasch Partial Credit Model.

(RES) RESAMPLING & SIMULATION TECHNIQUES

Poster 90: Correlation Coefficients in Cosine Space

Alexandria Ree Hadd, Vanderbilt University; Joseph Lee Rodgers, Vanderbilt University

The correlation coefficient be can interpreted as the cosine of the angle between centered variable vectors. Using this interpretation of the correlation, we re-draw the correlation space occupied by 3x3 correlation matrices first demonstrated by Rousseeuw and Molenberghs (1994). Once the cosine transformation is imposed on the space, the space occupied by 3x3 correlation matrices becomes a regular tetrahedron. A special type of 4x4 correlation matrices – banded correlation matrices – that can be represented in three dimensions is also re-drawn. Graphical demonstrations, extensions to larger correlation matrices, and implications for random correlation matrix generation are also discussed.

(SEM) STRUCTURAL EQUATION MODELING

Poster 91: Testing Indirect Effects in Meta-Analytic Structural Equation Modeling

Ivan Jacob Agaloos Pesigan, University of Macau; Shu Fai Cheung, University of Macau

Meta-analytic structural equation modeling (MASEM) combines the techniques of meta-analysis and structural equation modeling in order to synthesize correlation or covariance matrices from primary studies and fit structural equation models using the pooled matrix. The ability to test structural models in meta-analysis allows researchers to combine results from previous studies, to discover systematic patterns among the studies and to produce a more comprehensive summary of these studies. MASEM essentially involves two steps. First, it involves synthesizing correlation matrices from various studies to create a pooled correlation matrix. Second, the pool correlation matrix is fitted to SEM models (Cheung & Chan, 2005).

One situation in which MASEM can be very useful is mediation analysis. Using simulated and real data, the objective of this exposition is to examine the recent developments in testing the significance of indirect effect in ordinary least squares regression and SEM and to test whether the techniques for examining mediation effects in primary studies such as joint significance test of paths a and b, the delta method, bootstrapping, the Monte Carlo method and the distribution of products method, can be incorporated in MASEM. This endeavor benefits quantitative psychologists who are interested in developing methodologies and research psychologists who are interested in using appropriate methodologies to answer complex research questions both in the basic and applied fields.

Poster 92: Algorithm for Choosing the Number of Random Item-to-Parcel-Allocations in SEM

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

Parcel-allocation variability is variability in all parameter estimates, standard errors, and fit statistics across alternative allocations of items to parcels for a given factor. It arises even when items per factor are unidimensional in the population, with equal loadings (Sterba & MacCallum, 2010). Parceling remains widely used in SEM applications under conditions that lead to meaningful parcel-allocation variability (e.g., modest sample-sizes or communalities). Hence, researchers need to know how to quantify parcel-allocation variability within-sample. Of course, it would not be practical computationally to repeat the SEM analysis for every allocation from the finite population of millions of allocations possible using a given combination scheme of p_k parcels and q_jk items for parcel j=1…J_k of factor k=1…K. Instead, researchers may want to gauge how many randomly-drawn allocations, M, are required to obtain pooled parameter estimates and pooled standard errors (see Sterba & Rights, 2016) that are reasonably stable--even if a specified number of new allocations from that combination scheme are analyzed. Theory predicts that M should differ depending on particular model and data conditions (Sterba & Rights, 2016). This study presents an iterative algorithm for achieving this objective and describes alternative convergence criteria. This algorithm is implemented in a new R function, PoolMAlloc. Required input includes an item-level dataset, combination scheme, chosen convergence criteria, and parcel-level SEM specification. Output includes M and SEM results obtained using M allocations. This software is demonstrated with an empirical example. Advantages of an iterative algorithm over an arbitrary rule-of-thumb for choosing M are discussed.

Poster 93: Exploratory Structural Equation Modeling of Categorical Data

Jordan Prendez, University of Maryland

Recently, Exploratory Structural Equation Modeling (ESEM) has brought together the advantages of SEM with the flexibility of exploratory factor analysis. However, relatively little work has been conducted in order to determine ESEM performance when estimates are made from categorical data. In applied settings, it is often the case that categorical data (e.g., 5-point Likert data, binary correct-incorrect) are treated as continuous with the use of techniques that do not take into account the discrete nature of the data. Because treating categorical data as continuous leads to biased parameter estimates and standard errors several different approaches have been created for fitting latent variable models to categorical data. A recent comparison, specifically, found that categorical least-squares demonstrates low amounts of parameter bias and robust standard errors when compared to other methods when the number of discrete categories are fewer than 5 (Rhumtulla, Brosseau-Liard, & Savalei, 2012). This study aims to investigate the analysis of latent factors with categorical data from the relatively new ESEM framework and compare it to existing work conducted under a traditional SEM perspective. Conditions outlined in Rhumtulla et al., 2012 are used to compare bias in standard errors and parameter estimates from a categorical least-squares SEM framework to an exploratory SEM estimation framework. Results will be discussed within.

Poster 94: How Test Instructions Impact Motivation and Anxiety in Low-Stakes Settings

Catherine Mathers, James Madison University; Sara Finney, James Madison University; Aaron Myers, James Madison University

It has been empirically demonstrated in low-stakes testing contexts (e.g., TIMSS, institutional accountability testing) that perceived test importance affects test performance via test-taking effort (e.g., Zilberberg, 2013). However, other examinee characteristics such as test anxiety also impact test performance in low-stakes contexts (e.g., Hopko, Crittendon, Grant & Wilson, 2005; Sarason, 1961). Results from these studies show that increases in test anxiety lead to decreases in performance. Importantly, test anxiety has been found to increase as the relevance of the test to examinees increases in low-stakes testing contexts (Pekrun, Cusack, Murayama, Elliot &Thomas, 2014). Thus, our study examined how changing test relevance affects the relationships among test anxiety, performance, perceived test importance, and examinee effort. We hypothesized increasing the relevance of the low-stakes test moderates the mediated effect of perceived test importance on performance via examinee effort. Using data from a large-scale, low-stakes test completed by first-year college students for institutional accountability purposes (N = 1109), we manipulated test instructions to increase the relevance of the test to examinees. Results indicated no difference in average performance, test anxiety, examinee effort, and perceived test importance across the three test instruction conditions. Moreover, path analysis indicated the mediated effect of importance on performance via effort was not moderated by test relevance. However, as test relevance increased, anxiety had a greater effect on both effort and performance, thus increasing construct irrelevant variance and complicating test score interpretation.

Poster 95: Accounting for Careless Responding with Bogus Items and Response Time

Rung-Ching Tsai, National Taiwan Normal University; Ke-Chain Lin, National Taiwan Normal University; Tsui-Shan Lu, National Taiwan Normal University

A valid inference for a survey is always based on efficient estimation. The efficient estimation is threatened and distorted by careless responses. A bogus item is commonly used in survey questionnaire to detect the careless respondents and powerfully unveils those responding without reading the items. Response time is another way to identify the careless responses and its effect in enhancing item parameter estimation has been verified in educational testing. Here we propose a mixture response time model to analyze survey questionnaire with a bogus item. By distinguishing between two types of careless responding behavior using the bogus item as well as taking into account the response time, our approach aims to make better inference based on responses given by the attentive respondents. In the simulation studies we find that, for samples of size 500 or greater, the standard errors of the item parameters decrease as response time becomes available.

We further apply our proposed model to analyze a real data set with a large proportion of careless responding (22%) to illustrate the benefit of accounting for different careless responding behavior.

(SML) STATISTICAL & MACHINE LEARNING

Poster 96: Developing a New Off-Topic Advisory Flag in Automated Essay Scoring

Jing Chen, Educational Testing Service; Mo Zhang, Educational Testing Service

E-rater® is the automated scoring engine used at ETS to evaluate and score essays. A pre-screening filtering system is embedded in e-rater to detect and exclude essays that are not suitable to be scored by e-rater. The pre-screening filtering system is composed of a set of advisory flags, each of which marks some unusualness of the essay such as off-topicness. The advisory flags that detect off-topic essays usually requires topic-specific training essays in order to identify essays that are very different from the other essays of the same topic. In real test settings, topic-specific training essays may not available in time when new prompts are administered. To enhance the capability of off-topic advisory flags, in a previous study, we identified a set of features that were potentially useful in distinguishing off-topic essays that did not require topic-specific training essays. These features included essay length, essay organization, sentence variety, word variety, similarity of an essay to essays that classified by human raters in each score category. In this study, we built a new advisory flag using these identified features to predict the probability of an essay being off-topic. Results suggested the new advisory flag that utilize multiple essay features simultaneously identifies off-topic essays more accurately than the current advisory flags that consider one essay feature at a time.

(VAL) VALIDITY & RELIABILITY

Poster 97: Neglected Validities: The State of Validity Evidence in Preschool Assessment

Kathy Buek, University of Pennsylvania; Katie Barghaus, University of Pennsylvania

Quality, scientifically based assessment of major domains of children's functioning is an essential component of early childhood education. However, research has demonstrated that many of these assessments are developed and implemented without due consideration for fundamental aspects of their validity (Cizek, Rosenberg & Koons, 2008). The American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) developed rigorous standards by which to evaluate the evidence supporting the validity of inferences made from the results of an assessment or test (AERA, APA & NCME, 2014). According to the Standards, five types of validity evidence should be reported: evidence related to test content, response processes, internal structure, relationships to other variables, and consequences of test use.

The research team for the present study conducted a systematic review of all preschool assessments included in the Mental Measurements Yearbook published by the Buros Center for Testing. The study examines the types of validity evidence reported overall for these assessments as well as by test purpose and domain of child functioning. The results indicate trends and gaps in validity evidence reported for early childhood assessments. The study identified certain types of validity evidence that are often neglected in the process of developing, validating, and reviewing early childhood measures. The implications of these "neglected validities" are discussed. This presentation is designed to provide useful insights for test developers and reviewers, as well as for practitioners and researchers who administer early childhood assessments.

Poster 98: A Multitrait-Multimethod Factor Mixture Model to Assess Trait × Method Interactions

Kaylee Litson, Utah State University; Christian Geiser, Utah State University; G. Leonard Burns, Washington State University; Mateu Servera, University of the Balearic Islands

Multitrait-multimethod (MTMM) analyses are used to assess convergent and discriminant validity and to study method effects. Most current MTMM approaches assume that measures have equal convergent and discriminant validity across the entire range of trait values and thus do not account for potential trait × method interactions. In practice, convergent and discriminant validities might differ across unknown subgroups due to individual trait level differences (e.g., a high anxiety subgroup may show higher convergent validity for self- and other reports than a low anxiety subgroup). A novel approach is presented that allows analyzing trait × method interactions using factor mixture modeling (e.g., Muthén, 2001). The MTMM mixture model allows identifying latent classes of individuals who differ with respect to latent mean levels (i.e., traits) as well as convergent and discriminant validity (i.e., method effects). The new approach was applied to mother's and father's ratings of children's hyperactivity/impulsivity (HI) and inattention (IN; N = 618). Results revealed four latent classes: a "no symptoms" class, a "no HI/low IN symptoms" class, a "low symptoms" class, and a "moderate symptoms" class. The no symptoms and low symptoms classes showed strong evidence of convergent and discriminant validity whereas the moderate symptoms class showed weaker evidence for convergent validity and the no HI/low IN symptoms class lacked convergent validity for ratings of IN. Both mean levels and convergent validity estimates differed across latent classes, indicating the presence of a trait × method interaction. Advantages and limitations of the approach are discussed.

Poster 99: Meta-Analysis of Omega Composite Reliability: An Overestimation Problem Revealed

Rory M. Waisman, University of Manitoba; Johnson C.H. Li, University of Manitoba

Meta-analysis of a scale's reliability depends on the availability of reliability coefficients in published studies. The most widely reported reliability estimate, coefficient alpha (Cronbach, 1951), relies on assumptions that are frequently violated in practice and it tends to underestimate true reliability (Sijtsma, 2009). In recent years, some researchers have recommended the reporting of McDonald's (1999) omega-total, an estimate of reliability that is more robust to violation of the assumptions required for alpha (Dunn, 2014; Revelle & Zinbarg, 2009). With the availability of open-source software like R (R Development Core Team, 2013), computing omega-total is now easy and accessible. Thus, it is reasonable to expect that reporting of omega-total will increase. In light of this, the present simulation study evaluates the performance of omega-total under meta-analysis across 2,736 conditions. Results show that there is a tendency for omega-total to overestimate the true population omega-total. This problem is compounded by the fact that true population omega-total itself overestimates true reliability in some conditions (Simsek & Noyan, 2013). In general, the upward-bias of omega-total becomes more extreme and concerning when true reliability is low, sample sizes are low, general factor loading is high, and number of items is low. Our results reveal biases as high as 13%. Reliability is a critical psychometric property with implications for the interpretation of data collected using measurement scales. Just as underestimation of reliability by alpha has raised concerns, its overestimation prompts a cautionary note with respect to the reporting and interpretation of omega-total.

Poster 100: Person Fit Statistics Under Model Misspecification: A Monte Carlo Study

Ruben Castaneda, University of California-Merced; Nicole Zelinsky, University of California-Merced; Michelle Turitz, University of California-Merced

To interpret test scores, researchers employ various methods that demonstrate that the test has adequate psychometric properties. Sometimes, however, this may not be enough. Person fit statistics quantify the difference between expected item response patterns and observed item response patterns. Item response theory (IRT) assumes that the underlying trait is reflected by the response pattern. For response patterns consistent with the IRT model, the scores reflect the trait that is being measure. However, for inconsistent response patterns, the outcome score is unlikely to be meaningful or valid. Monte Carlo studies in person fit have focused on the statistics' abilities to capture aberrant response patterns. These studies have done so in the context of unidimensional IRT, but not multidimensional IRT to the authors' knowledge. Most psychological constructs are inherently multidimensional, but collapsed to unidimensional for application (e.g. scaling a personality measurement using unidimensional 2PL). This research focuses on identifying the outcome of Drasgow, Levine and Williams' (1985) Zh person fit statistics under null conditions when the data-generating model is missspecified. We consider 3 levels of misspecification (1 factor, 2 factors, and 3 factors) along with 3 levels for the number of items (15, 25 and 50) and 2 levels for the sample size (200, 1000). Suggestions for applied researchers are given.