IMPS 2018

Columbia University · New York City, New York, USA · July 9-13, 2018

Abstracts

Posters

Poster Session: 7/10/2018, 630 PM - 8:30 PM

Poster 1: A PSYCHOMETRIC INVESTIGATION OF A NONVERBAL ACCURACY ASSESSMENT
Beyza Aksu Dunya, University of Illinois at Chicago; Clark McKown, Rush University; Everett Smith, University of Illinois at Chicago

The Nonverbal Accuracy Assessment (NVA) aims to quantify a child's nonverbal signal threshold, defined as the lowest intensity nonverbal behavior that the child can accurately label (McKown et. al., 2009). The purpose of this study is to evaluate the psychometric properties of a Nonverbal accuracy assessment (NVA) designed for children in kindergarten through third grade. Data were collected from two separate and representative samples of children. The first sample included 4463 children and the second sample included 3218 children. The instrument includes 111 dichotomous items (0= Incorrect, 0= Correct) and the items contain images of male and female faces representing four emotions with differing intensities; joy, sadness, anger, and fear. Data were calibrated using Rasch dichotomous model (Rasch, 1960). We evaluated content, substantive, structural, responsiveness and generalizability aspects of construct validity. Differential item and test functioning were also evaluated across gender and ethnicity groups. Across both samples, consistent item fit, unidimensional item structure, and adequate item targeting were found. Poor item targeting at high measures indicated that more items are needed to distinguish high ability children. Analyses of DIF found some significant item-level DIF across gender, but no DIF across ethnicity. The statistical and graphical analyses of person measure calibrations with and without DIF items evidenced negligible differential test functioning (DTF) across gender and ethnicity groups in both samples. The results evidenced construct validity supporting use of the 111-item NVA assessment with different samples from the same target population. Adding harder items should improve targeting and reliability of the instrument.

(APP) Applications

Poster 2: EFFECTS OF OCCUPATIONAL STRESS ON SUBJECTIVE WELL-BEING AMONG PRIMARY TEACHERS
Meng Du, Collaborative Innovation Center of Assessment towards Basic Education Quality, Beijing Normal University; Tao Yang, Collaborative Innovation Center of Assessment towards Basic Education Quality, Beijing Normal University; Ruiyan Gao, Beijing Normal University; Xiaojian Sun, Beijing Normal University

This study aimed at exploring the effects of job burnout, stress, and salary on subjective well-being among primary teachers. A total of 7,657 primary mathematics teachers in East, Middle and West part of China completed a set of questionnaires which included the Subjective Well-Being Questionnaire (SWB), the Maslach Burnout Inventory Questionnaire (MBI), the Teacher Occupational Stress Factor Questionnaire (TOSFQ). The data was analysed by using the Structural Equation Modelling (SEM), the major findings of the study were as follows. First, primary teachers' stress had a negative effect on subjective well-being, after controlling the types of primary schools, teaching duration, sex, social-economic status and teachers' qualifications. Second, primary teachers' job burnout mediated the relationship between stress and subjective well-being. Third, teachers' salaries moderated the negative effect of occupational stress on subjective well-being. As the occupational stress increased, the subjective well-being of the teachers with a high salary was higher than those teachers with a low salary. The findings of this study suggest that the school leaders and the government need to pay more attention to the occupational stress of primary teachers, and provide better salaries for primary teachers who are in a high stress level in order to improve their subjective well-being.

(APP) Applications

Poster 3: THE EFFECT OF ITEM EXCLUSION FROM SCORING ON SCALE STABILITY
Yu Fang, ACT, Inc.; Yang Lu, ACT, Inc.

In testing programs, although not desirable, the number of valid items or score points to be reliably used for scoring may slightly change across forms occasionally. For example, after the test administration but before the score reporting, it was found that there was no correct key for one item

due to typo or misprint, or the biserial correlations for some items were substantially negative. In another example, one passage from the operational form was found breached before the testing, and the creation of an alternate form could be costly and time consuming. Both scenarios may result in the exclusion of some items, more or less, from the operational scoring. When items are excluded from the operational scoring, 'equating' still needs to be conducted to make scores comparable for reporting. The purpose of the present study is to investigate methods and conditions to better achieve the scale stability when items are excluded. More specifically, the goals of this study are to 1)

(APP) Applications

Poster 4: MODELING POLYSUBSTANCE USE USING THE NOMINAL RESPONSE MODEL
A. R. Georgeson, University of North Carolina at Chapel Hill

We modeled substance use items from a large national survey using the nominal item response model to investigate whether polysubstance use items could be scored on a single dimension. Data were from National Survey on Drug Use and Health (NSDUH). Items included the most commonly used substances, as well as classes of substances identified in the ICD-10. These included cigarettes, alcohol, marijuana, cocaine, LSD, ecstasy, prescription analgesics, prescription tranquilizers, prescription sedatives, and prescription stimulants. Items asked how recently the respondent had used each substance and responses included "Within the past 30 days", "more than 30 days ago but within the past 12 months", "more than 12 months ago", and "never used". A unidimensional model was fit to the data and local dependence statistics were used to guide the specification of additional multidimensional models. Results supported a two-tier model with all items loading on a general factor and clusters of items loading on group-specific factors. The explained common variance (ECV) statistic was used to summarize the proportion of variance explained by a general factor. This index was used to determine whether a general polysubstance use factor accounted for enough variance to support scoring use of a single score. To our knowledge, there is no continuous measure of polysubstance use in the literature and this application of the nominal response model provides support for a general substance use dimension. Scoring possibilities based on the nominal model score function values are discussed, as well as the substantive interpretation and use of the scale.

(APP) Applications

Poster 5: SPECIFYING "EXPERT" RATERS: AN APPLICATION OF THE RATER ACCURACY MODEL
Madison Holzman, James Madison University; Aaron Myers, James Madison University; Allison Ames, James Madison University

Performance assessments (PAs) typically require human raters. Expert rater scores are often obtained from content experts, and their scores are lauded as the "gold standard" and used as evidence for the meaning of scores. However, due to a limited number of content experts and a vast number of PAs, it is infeasible for content experts to provide scores for all PAs. Thus, operational raters are trained in hopes that they will mimic content experts' scores. Considering content experts spend years learning information related to their topic, is it realistic to expect that operational raters can effectively produce scores similar to those of content expert raters? Or, does an experienced operational rater serve as an effective proxy that is more realistic for operational raters to mimic? Authors propose operationalizing an "alternative expert rater" as a highly experienced rater. The primary research question is whether, in the context of ethical reasoning, rater accuracy differs if scores are compared to a context expert or alternative expert rater. In this study, data were collected from 1) twelve operational raters trained on the rubric, 2) a content expert rater, and 3) an alternative expert rater who had participated in more than five essay rating sessions. To answer the research question, the Rater Accuracy Model (RAM; Engelhard, 1996) is extended to include an interaction between expert rater and rater accuracy. Results may have implications for the selection of expert raters. Implications for the validity of scores will be discussed.

(APP) Applications

Poster 6: EFFECTS OF ANCHOR ITEM PARAMETER ON TEST EQUATING
Keonseob Kim, Yonsei University; In-Yong Park, Korea Institute for Curriculum and Evaluation; Guemin Lee, Yonsei University

This study will investigate the degree to which IRT equating is affected by difference of item parameter of anchor test in the common-item non-equivalent groups design. Condition is (a) difficulty, (b) discrimination, and (c) the ratio of questions that differs among the total anchor test.

(APP) Applications

Poster 7: APPROPRIATE RELIABILITY COEFFICIENTS DEALING WITH MULTIPLE–YET SINGLE ESSAY PER EACH-PROMPTS
Sunhee Kim, The College Board; Michael E. Walker, The College Board; Xiuyuan Zhang, The College Board; Weiwei Cui, The College Board

Traditional reliability coefficients (e.g., Cronbach's alpha and weighted Kappa) require at least two items or essays from the same type of task. When examinees respond to multiple essay prompts yet only one essay from each task types (e.g., an examinee writes one essay on Issue prompt and another essay from Argument prompt), such analyses may not be suitable—due to differences in what the tasks are measuring. Furthermore, traditional reliability measures could be confounded by: (a) the correlation between different tasks; and (b) the consistency of rater performance. This simulation study applied multiple reliability measures on test scores from multiple tasks where each task comprises one essay. We examined FACETS model (Linacre, 1997) and G-theory (Brennan, 2001) approaches to dissect the impact of task correlations and rater performance on the reliability given a complex essays-within-tasks nesting structure. The simulation data was generated by a hierarchical rater model with signal detection theory (HRM-SDT; DeCarlo, Kim, & Johnson, 2011), which allows various conditions on rater performance and the true score correlations between task types while representing nested nature of data (e.g., raters are nested in essay & essays are nested in task).As a result, FACETS Rasch test score reliability estimates were high (possibly inflated) and carrying rater performance; while the G-Theory true score reliability performed better on controlling rater effects. Both Rasch and G-Theory rater reliabilities represented rater performance, but Rasch rater reliability was unstable on homogeneous raters. Extended simulations will be included and implications for practice will be discussed.

(APP) Applications

Poster 8: WHAT ARE THE IMPACTS OF ASSESSMENTS ON AN ADAPTIVE LEARNING PLATFORM?
Bor-Chen Kuo, National Taichuang University of Education

On April 2017, the Ministry of Education, Taiwan, launched an adaptive learning and assessment platform (http://adaptive-learning.moe.edu.tw/) to assist student learning. Until now there are more than 60,000 users. In this platform, there are thousands of videos for micro-learning, items for instant diagnosing, interactive modules and intelligent tutoring agents for supporting learning through scaffolds.As the main development team, we were very interested in how did the different types of assessments on this platform can help student's learning or classroom teaching (flipped classroom). In this platform, we designed in-video quiz, after-video practice item, dynamic assessment, conversation-based assessment and unit-based adaptive diagnostic assessment. In this presentation, we try to explore: 1. Can the in-video quiz help student learning? Does it change the video-watching behavior? 2. Does the adaptive assessment help student learning in after class remediation? How effective compare to traditional paper-pencil after class remediation. 3. Additionally, student's learning behaviors and records are analyzed by using data mining techniques and the findings will be also presented.

(APP) Applications

Poster 9: SHINYITEMANALYSIS FOR PSYCHOMETRIC TRAINING AND RESEARCH
Patricia Martinkova, Faculty of Education, Charles University, and Institute of Computer Science of the Czech Academy of Sciences; Adéla Drabinová, Faculty of Mathematics and Physics, Charles University, and Institute of Computer Science of the Czech Academy of Sciences

ShinyItemAnalysis, R package and online application (Martinková, Drabinová, et al., 2017), has been developed for teaching psychometric concepts, presenting psychometric research and to foster use of psychometric methods in educational test development. In this work we present useful features of ShinyItemAnalysis such as presence of real and simulated data examples, model equations, interactive plots, parameter estimates, interpretation of results, selected R code, or automatic report generation. Further, we describe how ShinyItemAnalysis helps to promote psychometric research in classical test theory (Martinková, Štěpánek et al., 2017) or in detection of differential item functioning (DIF) by providing real and simulated examples showing that DIF occurs independently of difference in total scores (Martinková, Drabinová, Liaw, et al., 2017) and by providing newly developed DIF detection methods (Drabinová & Martinková, 2017). Finally, with an example of admission test to medical school, we show how ShinyItemAnalysis may provide a free and user friendly tool to routinely analyze tests and to foster use of psychometric methods in educational test development.

(APP) Applications

Poster 10: APPLYING MIRT TO THE VALIDATION AND STANDARDIZATION OF A TURNOVER INTENTIONS SCALE
Igor Menezes, University of Lincoln; Ana Cristina Menezes, Federal University of Bahia; Kai Ruggeri, Columbia University; Patryk Muszynski, University of Lincoln

The purpose of this study was to make available to researchers and practitioners a new instrument to measure turnover intentions based on a compensatory Multidimensional Item Response Theory (MIRT) model and to introduce an alternative procedure for test standardization (Weighted Scores) under a MIRT approach. The Multidimensional Turnover Intentions Scale (MTIS) was administered to 146 workers of a multinational automotive company. Since extrinsic and intrinsic aspects were measured, item parameters and individual scores are provided for each dimension. Multidimensional Graded Response Model was chosen for item calibration and EAP estimation technique was deployed for producing the individual's factor scores. The two-dimensional structure was confirmed, with all the thirty items properly measuring turnover intentions. Items more likely to predict turnover intentions and an interpretation about individual scoring under a MIRT approach are presented. Finally, Weighted Scores were calculated based on the individual responses and the discrimination for each dimension. When compared to estimated factor scores, they show very attractive psychometric properties, with no items exhibiting differential item functioning, and no significant differences for the total test information and expected total score distributions. This suggests that Weighted Scores could be used in place of factor scores under a MIRT approach, mainly when paper-and-pencil instruments are scored manually or when more advanced procedures for the estimation of factor scores are not available to test administrators. The MTIS can help companies to work on the analyses of their talented employees with stronger intentions to leave, and then create strategies aimed at their retention.

(APP) Applications

Poster 11: IRT ANALYSIS OF THE STUDENT ATTITUDES TOWARDS STATISTICS-36 SCALE
Brooke Midkiff, University of North Carolina at Chapel Hill

The Student Attitudes Towards Statistics-36 scale (SATS-36) contains 36 items that measure college students' overall attitudes towards statistics via six factors: affect, cognitive competence, value, difficulty, interest, and effort (Schau, Stevens, Dauphine, & Del Vecchio, 1995). The scale has been previously analyzed using confirmatory factor analysis (CFA) (Vanhoof et al., 2011). No previous studies have examined item functioning or model fit for known groups other than gender, and no research to date has utilized Item Response Theory (IRT) to examine the overall factor structure, possible Differential Item Functioning (DIF), or Local Dependence (LD). This research fills this gap in the literature by providing an IRT analysis of item functioning and the overall factor structure of the SATS-36. The sample

consists of undergraduate students from three different sections of an introductory statistics course in Psychology and Neuroscience over the course of two semesters. Samejima's (2010) graded response model was fit for each subscale as separate, unidimensional models. A multidimensional graded response model combining the subscales was also fit, and then, following the recommendations of Vanhoof et al. (2011), iteratively refit. Analyses were conducted using IRTPRO 4 (Cai, Thissen, & DuToit, 2017). DIF was examined by gender, first-generation college student status (FGCS), underrepresented minority status (URM), student year in college, and transfer-student status. LD was examined within each unidimensional model as well as within the multidimensional model. Findings indicate significant LD and DIF among several subscales and between multiple known groups. Suggestions for scale refinement based on the IRT analysis are provided.

(APP) Applications

Poster 12: QUANTIFY THE EVIDENCE OF THE NULL EFFECT OF DISFLUENCY

Shunta Nagano, Yamaguchi University; Koji Kosugi, Yamaguchi University; Fuminori Ono, Yamaguchi University

We examined whether disfluent font is more effective for learning or not and estimated an effect size of disfluent font on learning. Studies investigating the effect of disfluent materials on learning outcomes (i.e. the effect of disfluency) have generated controversy whether it exists or not (c.f. Kühl & Eitel, 2016). However, previous studies failed to quantify the evidence of null hypothesis since p-value does not represent the evidence of null hypothesis. We resolve this problem by Bayesian approach which is able to quantify the evidence of null hypothesis as well as alternative hypothesis. One hundred and thirty-nine undergraduates participated in an experiment (male 74, female 65). Our experimental procedure follows Diemand-Yauman et al. (2011)'s study1. We randomly divided the participants into two groups learning with disfluent font or fluent font. Participants took a delayed recall task after learning. We constructed a hierarchical model which is structured differences in individual and items. In consequence, Bayes Factor is about 6-9 times in favor of null hypothesis rather than a hypothesis assuming the positive disfluency effect. Furthermore, expectation value of the posterior distribution of the effect size is -0.07, and 95 percent credible interval of that is from -1.01 to 0.86. In addition, it is about 13% that estimated effect size is higher than effect size reported in Diemand-Yauman et al. (2011)'s study2. We conclude that the effect of disfluency on learning does not have certain evidence to implement in the real-world classroom.

(APP) Applications

Poster 13: ADDRESSING DATA NON-NORMALITY: CURRENT TRENDS AND FUTURE RECOMMENDATIONS

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

Data non-normality is among the most common experiences encountered in statistical practice (e.g., see Cain, Zhang, & Yuan, 2016; Micceri, 1989). Because the linear model often serves as the basis for analyzing data, we review extant developments on how best to address non-normality with the linear model. Example approaches are popular data transformations (Box & Cox, 1964; Tukey, 1977) and reverse transformations (e.g., Duan, 1983), trimming (Wilcox, 2017), and the use of robust estimators (e.g., White, 1980). In particular, we highlight important philosophical distinctions and diverse downstream outcomes when these methods are applied to non-normal data. Several empirical examples will be used to illustrate important distinctions between these alternative methods. A nuanced appreciation of the underlying motivations behind competing methods to address data non-normality is pertinent to formulating theories, appropriately analyzing data, and accurately representing complex psychological data.

(APP) Applications

Poster 14: A MODIFICATION OF IRT-BASED STANDARD SETTING METHOD
Pilar Rodriguez, University Center East Regional, University of the Republic, Uruguay

We present a modification of the IRT-based standard setting method proposed by Garcia, Abad, Olea & Aguado (2013) that we combine with the Cloud Delphi method (Yan, Zeng & Zhang, 2012). García et al. (2013) calculated the average characteristic curve of each level. In the proposed new method, the influence of each item on the average is weighted according to its proximity to the next category. This method improves the cut scores estimation. Performance levels are placed on a continuous scale, asking each judge to score each item on the scale. The Cloud Delphi method is used until a stable final score is achieved. From these judgments, the weights of each item in each level are calculated. Then, for each item family that indicates a certain performance level, the weighted average characteristic curve is calculated. In the next step, joint averaged-ICC is calculated. We define the probability that an examinee with the minimum required knowledge will correctly answer the item to calculate the cut scores of level k. The corresponding theta is calculated, this being the cut scores of the level k. In this work, the modified method is applied for different probability values and these results are compared with Bookmarking method.

(APP) Applications

Poster 15: EFFECT OF TREND SET DESIGN ON DETECTING HUMAN RATER ERROR
Adrienne Sgammato, Educational Testing Service; John R. Donoghue, Educational Testing Service

Open ended items are commonly used on educational assessments. Such items are usually scored by human raters by assigning a score based on the item's scoring guide. For assessments that report on performance trends over time, it is important to ensure that items are scored in the same way across test administrations. Various approaches exist for doing this including interspersing a subset of Time 1 papers into Time 2 scoring, calculating and evaluating reliability/agreement statistics such as t, kappa, percentage of exact agreement. Criteria for selecting the subset of trend papers (trend sets) is important to be sure that all score points are well represented and also that there are sufficient papers to capture various sources of rater disagreement. Given the various errors that raters can make in assigning scores (e.g., random, rater centrality, too lenient/strict, category confusion), this study was designed to evaluate the effect of trend set design on ability to detect rater errors. Using a fully factorial design, 6 levels of rater error, 4 levels of trend set designs (strictly uniform; strictly proportional; variants of those), 6 levels of trend set size (100, 200, 400, 600, 800, 1600), and 4 levels of item score categories (3, 4, 5, 6 category items) were defined. Rater agreement statistics between Time 1 and Time 2 include: t, unweighted Cohen's kappa, Stuart's (1955) Q, and percentage of exact agreement. Results will indicate, of those defined here, the best trend design for each kind of rater error.

(APP) Applications

Poster 16: EVALUATION OF MORAL DISENGAGEMENT SCALE BASING ON ITEM RESPONSE THEORY
Yanni Shen, Beijing Normal University; Xiaojian Sun, Beijing Normal University; Liping Yang, Beijing Normal University; Tao Xin, Beijing Normal University

Moral disengagement refers to social cognitive mechanisms that enable individuals to behave immorally for the sake of self-interest. The moral disengagement scale, which was developed by Bandura, Barbaranelli, Caprara, and Pastorelli (1996), has been widely used in many countries and proved to have good reliability and validity in previous research. This study employed item response theory (IRT) analyses to investigate the psychometric properties of the moral disengagement scale using a Chinese sample. Exploratory factor analysis identified a single latent trait. The grade response model (GRM) fitted the data better comparing with partial credit model (PCM) and generalized partial credit model (GPCM). The item-level fit indices showed that none significance were found for the scale items after Bonferroni adjustment. In addition, all items were found to have moderate to high discrimination values (ranged from 0.59 to 2.76), which indicates that the scale has a good capability of discriminating different levels of moral disengagement. Intervals of the four corresponding threshold parameters are [−3.46, 0.76], [−1.40, 1.52], [0.21, 2.86], [2.35, 4.79], respectively, which indicate that individuals tend to choose grades at middle ranges. Results suggest that the psychometric properties of the moral

disengagement scale are desirable. The moral disengagement scale can be further adopted for Chinese samples.

(APP) Applications

Poster 17: CAN ANXIETY AND DEPRESSION BIAS COGNITIVE PERFORMANCE?
Felipe Valentini, University Salgado de Oliveira; Heungsun Hwang, McGill University

This study aims to investigate the impact of humor on the item parameters of an intelligence test from different perspectives, using latent variable models. Nine hundred ninety-eight Brazilian students were administered the Abstract and Spatial Reasoning Intelligence Test, Baptista's Depression Inventory, and Questionnaire of Anxiety. We considered three different latent variable models (MIMIC, multiple groups, and factor mixture). In all tests, depression was not related to either intelligence scores or item parameters, so that it was removed from the models. In the MIMIC, anxiety was linearly related to intelligence scores, whereas it was slightly related to five item parameters only, suggesting a trivial bias. For the multiple group analysis, we classified the participants into five groups according to their anxiety scores and tested if these groups preserved the invariance of the item parameters. Scalar invariance across the five groups was supported, indicating that the item parameters were unbiased. We finally used mixture models to examine if anxiety might predict any intelligence latent class. A model involving one latent factor with two latent classes was fitted to the data. Anxiety was associated with the latent classes, each of which showed a different level of the participants' scores. However, the classes involved only negligible item parameter differences. Thus, anxiety might be related to intelligence performance, but only through the latent scores, whereas it did not bias the item parameters. This indicates that the intelligence test may be used for accessing cognitive performance of the participants even with different levels of humor.

(APP) Applications

Poster 18: INVESTIGATING ITEM-POSITION AND PSYCHOLOGICAL FACTOR EFFECTS ON ITEM PARAMETER ESTIMATION
Nayeon Yoo, Teachers College, Columbia University; Young-Sun Lee, Teachers College, Columbia University

Item-position effect is defined as the impact of the position of an item within a test on item characteristics. Prior studies reported the effect of item positions on item parameters and equating results when item positions differ across several administrations or forms. Recent works suggest that item-position effects could vary depending on factors such as direction/degree of position change, or test taker's ability level. Psychological factors could be another important factor; we aimed to examine the effects of item positions and psychological factors on item parameter estimates under Item Response Theory (IRT) framework via structural equation modeling (SEM). Real-world data analyses were conducted using TIMSS 2015 Grade 8 Mathematics data, and simulation studies with psychological factor conditions were conducted to examine the recovery of parameters. TIMSS employs a matrix-sampling method, grouping the items into a number of item blocks. Each booklet has its own order of item blocks, each item block appearing in two booklets in different locations. In this study, item responses of block 01 at the beginning and the end were analyzed. To examine the effects of item positions and psychological factors on item parameter estimation, first, DIF analysis was conducted in order to detect any existing item-position DIF. Then, propensity score matching was employed to form matched sets of subjects who had items at the beginning and at the end, having similar ability level as the propensity score. Finally, relationships between calibrated item difficulties under 2PL and psychological responses were explored via SEM.

(APP) Applications

Poster 19: DIAGNOSING INNOVATIVE AND MEASUREMENT OUTLIERS IN MULTI-SUBJECT TIME SERIES DATA

Dongjun You, Pennsylvania State University; Sy-Miin Chow, Pennsylvania State University; Michael Hunter, University of Oklahoma; Meng Chen, University of Oklahoma

Outliers in times series data can be more problematic than in independent observations due to the correlated nature of such data. It is common practice to discard outliers as they are typically regarded as a nuisance or an aberrance in the data. However, outliers can also convey meaningful information concerning potential model misspecification, and ways to modify and improve the model. Moreover, outliers that occur among the latent variables (innovative outliers) have distinct characteristics compared to those impacting the observed variables (measurement outliers), and are best evaluated with different test statistics and detection procedures. We provide illustrative examples to showcase dynr.taste - a function for detecting outliers in state-space models proposed by de Jong, and Penzer (1998) for single-subject time series data and later extended by Chow, Hamaker, and Allaire (2009) to multi-subject settings. Results from a Monte Carlo simulation study are used to shed light on factors to consider in performing outlier detection in multi-subject time series data. We demonstrate the empirical utility of the proposed approach and software function using data from an ecological momentary assessment study of emotion regulation.

(APP) Applications

Poster 20: EFFECTS OF PARENT-TEACHER RELATIONSHIP ON STUDENTS' ACADEMIC ACHIEVEMENT IN CHINA

Ying Yuan, Beijing Normal University; Guan-Yu Chen, Beijing Normal University

Although many researches have shown that home-school cooperation has an important impact on students' academic achievement, the parental-teacher relationship in home-school cooperation has escaped the attention of many Chinese researchers, especially not prevalence in quantitative research. In order to study the complex impact of parent-teacher relationship on students' academic achievement, this study which is based on the China Education Panel Survey（CEPS） uses the Hierarchical Linear Models (HLM) to empirically find the factors affecting students' academic achievement, especially highlights the influence of the parent-teacher relationship on junior middle school students' achievement. Besides, this study uses Propensity Score Matching (PSM) to find the net effect of the parent-teacher relationship to students' achievement. The preliminary study finds that parents' actual participation in school activities, parents' active contact with teachers and parents' social class had no significant influence on students' academic achievement. However, the willingness of parents to participate in activities, teachers' active contact with parents, whether parents are afraid of contacting teachers, and parents' education level influence students' achievement significantly. Meanwhile, this study further analysis how teachers' teaching age, education level, job satisfaction, perceived degree of parental respect, school's level, school district, and school resources to influence the parent-teacher relationship to affect students' academic achievement. These findings suggest that teachers should build a good communication and cooperation mechanism with parents and maintain an equal dialogue relationship. Parents should recognize their role in home-school cooperation, actively assume the responsibility and use corresponding rights, and jointly promote the healthy development of students.

(APP) Applications

Poster 21: SOCIAL MEDIA FATIGUE SCALE (SMFS): DEVELOPMENT AND VALIDATION

Shiyi Zhang, Collaborative Innovation Center of Assessment toward Basic Education Quality, BNU; Tao Xin, Beijing Normal University; Yilu Wang, School of Psychological and Cognitive Sciences, Peking University, Beijing; Xiaotong Zhang, King's College London

Objective: The present study aims at figuring out the construction of Social Media Fatigue (SMF) and developing a valid measurement: Social Media Fatigue Scale (SMFS). Methods: Based on the signs of SMF (Technopedia., 2011; Huff, 2014) and a pilot survey (n=30), the preliminary version of SMFS included 15 items. Other 52 items were used for validity analysis or collecting demographic variables. All

the items were rated by 7 grades (1 = Totally Disagree, 7 = Totally Agree). An Exploratory Item Analysis was implemented in R environment (using mirt package) for determining the dimensions of SMFS and selecting items (n=519). The item parameters, Concurrent Validity and Criterion-Related Validation of final scales were evaluated. Results Based on the analysis results, we finally choose the 3-dimension model to fit SMFS, and deleted 3 items to form the final version of SMFS (12-item SMFS). Item fit analysis showed that all the items fited well. The Cronbach's alpha of 12-item SMFS was 0.804, and IRT modeling revealed good reliability for 3 dimensions (r = 0.713 - 0.797), indicating that the final version has a acceptable reliability. We assessed Concurrent Validity against scores on participants' subjective feeling of SMF. And the correlations between SMF and other related variables (such as social media helpfulness, social media confidence, concerning about privacy, etc., Bright, Kleiser, & Grau, 2015) were computed to assess the Criterion-Related Validation. The significant correlations (r=0.272-0.483, ps<.01) reflected that the final scale has satisfactory Concurrent Validity and Criterion-Related Validation.

(APP) Applications

Poster 22: PARENTING AND STUDENTS' ACADEMIC ACHIEVEMENT: A MODERATED MEDIATION MODEL

Tingdan Zhang, Beijing Normal University

Filial piety is a Confucian practice that exhibits the close connection between Asian students and their parents which may contribute to their academic excellence (Yeh, 2003). A cross-sectional study was conducted. A total of 831 elementary students (M ± SD = 10.29±0.88 years; 46.0% females) were recruited from Hebei Province, China. The children completed a self-administered questionnaire that asked questions about their demographics, perceived parental education expectation, reciprocal and authoritarian filial piety. They brought home a questionnaire for their parents, including questions about parents' educational involvement and mother's and father's educational attainment. Measures of children's academic achievement were obtained from school records. We tested the moderated mediation model by multiple linear regression analyses, using the analytic procedures recommended by Muller et al. (2005). The model was tested using the bootstrapping in SPSS 22.0 (1000 bootstrap samples). After gender, age and parents' education level were included as covariates, the results revealed that perceived parental education expectation was significantly associated with academic achievement. Children's filial piety played a mediating role in the relationship between perceived parental expectation and academic achievement. Reciprocal authoritarian filial piety played a positive role while authoritarian filial piety played a negative role. Furthermore, maternal educational involvement moderated the mediated path from authoritarian filial piety through academic achievement. With maternal educational involvement increasing, the negative effect of authoritarian filial piety on academic achievement reduced. Therefore, the effect of perceived parental expectation on children's academic achievement was moderated mediating effect. The moderating effect didn't exist in paternal educational involvement, though.

(APP) Applications

Poster 23: NUTS FOR MIXTURE IRT MODELS

Rehab Al Hakmani, Southern Illinois University Carbondale

The fully Bayesian estimation via the use of Markov chain Monte Carlo (MCMC) techniques has become popular for estimating item response theory (IRT) models. The No-U-Turn Sampler (NUTS) is a relatively new MCMC algorithm that avoids the random walk behavior that common MCMC algorithms such as Gibbs sampling or Metropolis Hastings usually exhibit. Given the fact that NUTS can efficiently explore all dimensions of the target distribution, the sampler converges to high-dimensional target distributions more quickly than other MCMC algorithms (Hoffman & Gelman, 2014) and is hence less computational expensive. Previous research has applied NUTS to simple IRT models and demonstrated its advantage over Gibbs sampling in this aspect (Zhu, Robinson, & Torenvlied, 2014). However, to date, no research has adopted it to fit the more complex mixture IRT model, which is used under the situation where subpopulations perform differently on the same set of items. Therefore, the focus of this study is to apply NUTS to the two-parameter (2P) mixture IRT model. Monte Carlo simulations are carried out to investigate: 1) parameter recovery of the 2P mixture IRT model; 2) the prediction accuracy of class

memberships of individual persons from such a model; 3) the performance of the 2P mixture IRT model in comparison to the conventional 2P IRT model when latent classes exist. In addition, real data from tests that involve speediness are used to illustrate the comparison of the mixture IRT model against the conventional IRT model via the implantation of NUTS.

(BSI) Bayesian Statistical Inference

Poster 24: ESTIMATING ITEM PARAMETERS UNDER BI-FACTOR MIRT MODEL USING MCMC
Lida Chen, University of Iowa; Shumin Jing, University of Iowa

The purpose of this project is to examine the performance of using Markov Chain Monte Carlo (MCMC) to estimate the item parameters under the Bi-Factor MIRT model. A simulation study was conducted. Bi-Factor MIRT model was used for data generation and parameter estimation. Specifically, the software WinGen (Han, 2007) was used to draw item and examinee parameters, and flexMIRT (Cai, 2017) was used to generate response strings based on the parameters. A total of 1000 examinees, as well as their responses to the 30 test items, were simulated. Three sets of priors, including informative, vague, and non-informative priors, were selected for the purpose of item parameter estimation using OpenBUGS (Spiegelhalter et al., 2017). Both history plots and BGR diagnostic plots showed that although estimation all converged under three different sets of priors, many more iterations were required to reach convergence for item parameters under vague priors and non-informative priors, compared with informative priors. In addition, the statistic, RMSE, was used to evaluate the estimation precision across conditions. Results showed that using informative priors would generally make convergence faster and produce more precise item parameter estimation. This study demonstrated that MCMC is able to recover the item parameters under a Bi-Factor MIRT model at a satisfying level of precision when the priors had some degree of information.

(BSI) Bayesian Statistical Inference

Poster 25: ON SETTING THE SCALES OF LATENT VARIABLES IN BAYESIAN SEM
Benjamin Graves, University of Missouri; Edgar Merkle, University of Missouri

It is well known that, in traditional SEM applications, a scale must be set for each latent variable by setting either the latent variance or a factor loading to one. While this has no impact on the fit metrics in ML estimation, it can possibly lead to a preference in Bayesian model assessment metrics due to the use of different priors under each parameterization. Using a single-factor CFA as motivation for the study, we first show that Bayesian SEM fit metrics can be preferential toward one model or another depending on how priors are specified. We then propose a solution that involves prior distributions on ratios of model parameters. We illustrate the solution via simulation studies and an application to real data.

(BSI) Bayesian Statistical Inference

Poster 26: THE COUNT SOCIAL RELATIONS MODEL: TO BAYES OR NOT TO BAYES
Justine Loncke, University of Ghent

The family social relations model (SRM) is widely used to identify the sources of variance in interpersonal dispositions in families. Traditionally, it makes use of dyadic measurements that are obtained according to a round-robin design, where each family member rates the other family members on a specific interpersonal disposition. In this study, we employ the blocked design which is restricted to merely intergenerational dyadic measurements (e.g. parent-child). The parameters of the SRM can be estimated by defining it in a multilevel framework and applying a Bayesian method using Gibbs sampling. Such Bayesian approach has already been proposed for continuous outcomes for the SRM without family roles. Here, we aim to extend that approach to the SRM with family roles and to accommodate it to a count outcome variable. However, this approach might result in biased parameters of the variances for such a small group size in combination with a small sample size. Alternatively, the parameters of the model can also be estimated using the traditional SEM-framework by approaching the SRM-analysis as a confirmatory factor analysis (CFA) with count indicators. We perform a simulation study where the

performance in the Bayesian framework is compared to the performance of CFA with count factor indicators in the SEM-framework

(BSI) Bayesian Statistical Inference

Poster 27: COMPARISON OF FREQUENTIST AND BAYESIAN MODEL AVERAGES

Sinan Yavuz, University of Wisconsin, Madison; David Kaplan, University of Wisconsin, Madison

Model averaging approaches have been developed within both the frequentist and Bayesian paradigms of statistical inference. However, these approaches to model averaging have not been compared directly to each other in terms of their predictive accuracy. For frequentist model averaging (FMA), model selection is typically based on Akaike's information criterion. On the other hand, Bayesian model averaging (BMA) is based on Bayesian information criterion. The main goal of this study is to compare predictive accuracy of the two methods. However, to examine the contribution of the model averaging, frequentist linear regression and Bayesian linear regression methods are also added to the comparison. This research is conducted as a simulation to reduce possible noise and compare the methods objectively. In our design of the study, three different sample sizes (100, 500, 1000) and two different correlations among predictors (0.2, 0.6) are used. In total 6 different conditions are created. Each condition is replicated for 500 times. In each condition 10 multivariate normal variables are generated under the given conditions. Predictive performances is determined by calculating the predictive coverage and the log-score of predictive of predictive coverage. Results are averaged and presented over replicates to provide a frequentist evaluation of the FMA and BMA methods.

(BSI) Bayesian Statistical Inference

Poster 28: BALANCE AND TREATMENT EFFECT ESTIMATION IN THE PRESENCE OF COVARIATE MEASUREMENT ERROR

Heather Harris, James Madison University; S. Jeanne Horst, James Madison University

Propensity score matching techniques provide a means by which researchers can control for selection bias through the creation of a comparison group that is qualitatively similar to participants on key covariates (Austin, 2011; Stuart, 2010; Stuart & Rubin, 2008). However, in order to adequately implement PSM techniques, a researcher must decide which covariates to include in the model and how to measure each covariate (Guo & Fraser, 2014; Millimet, 2011; Rudolph & Stuart, 2017; Steiner, Cook, & Shadish, 2011). Best practices literature includes recommendations to administer instruments that result in reliable covariate scores (Guo & Fraser, 2014; Steiner et al., 2011). However, little is known about how numeric matching diagnostics perform when error-prone covariates are used to created matches. The accuracy of treatment effect estimates and the performance of numeric balance diagnostics were evaluated when error-prone covariates were included in the model. Data were simulated in R version 3.4.3 (R Core Team, 2017), and four propensity score matching methods were used to create qualitatively-similar treatment-comparison groups (nearest neighbor matching, nearest neighbor matching with a 0.2 caliper width, optimal matching, and Mahalanobis distance matching). Both error-prone (i.e., covariate scores simulated with measurement error) and error-free (i.e., covariate scores simulated without measurement error) covariate sets were employed. Type of measurement error and level of measurement error varied across twelve simulated conditions. Even when groups appeared balanced, as the level of measurement error increased, bias in the estimated treatment effect also increased. Implications for applied researchers are offered.

(CAU) Causal Inference and Mediation

Poster 29: PRIOR STATISTICAL KNOWLEDGE AND PERFORMANCE: MEDIATOR ANALYSIS OF STATISTICS ANXIETY

Oluwagbemisola Ladipo, University of Manitoba; Johnson Li, University of Manitoba

Research currently shows that statistics anxiety is a very important factor in predicting performance in statistics courses. This study expands on previous research by investigating the mediating effects of statistics anxiety on the relationship between students' prior statistical knowledge and their final grades.

A moderation analysis will also be conducted to examine whether demographic variables, namely age, gender and ethnicity, could further explain or moderate the mediation effects from statistics anxiety. Data collection is ongoing, and the final sample will consist of psychology undergraduate students from a major university in Central Canada. Participants will be asked to fill out a self-report survey containing questions from the Statistics Anxiety Rating Scale (STARS), demographic information such as age and gender, and questions investigating their previous statistical knowledge. The information gathered will be analyzed to examine if the effects of prior statistical knowledge has been heightened or lessened by statistics anxiety, and if this change of anxiety state differs across age, gender, and ethnicity. The results of this study will be used to guide and improve methods of teaching research methodology courses, and creating extra statistics workshops to alleviate the effects of statistics anxiety on final grades. In addition, if prior knowledge is shown to have a significant effect on final grades, it opens up a new method of improving grades of students by increasing their knowledge of statistics through optional classes for introductory statistical concepts before being graded on it.

(CAU) Causal Inference and Mediation

Poster 30: EXAMINING CONDITIONAL INDIRECT EFFECTS AT SAMPLE STATISTICS OF THE MODERATOR

Yu Liu, University of Houston; Jenn-Yun Tein, Arizona State University

In mediation analysis, if a mediational path is moderated, researchers typically choose a few conditional values on the moderator (say Z) at which to examine the significance of the conditional indirect effects. Often researchers choose the numeric value of a few sample statistics (e.g., the sample mean, and one sample standard deviation above and below the sample mean) of the moderator Z, and make inferences of the conditional mediated effect at the statistic (e.g., the mean) of the moderator Z rather than the corresponding numeric value (e.g., Z = 5.117). When Z is randomly sampled, the sample statistics of Z may not be identical from sample from sample. However, typical practices do not account for such sampling variability of the sample statistics of Z, and the resulting significance test of conditional indirect effects at sample statistics of the moderator potentially can be inaccurate. Using Monte Carlo simulation, this study investigates the Type 1 error rate and the coverage rate of 95% confidence intervals of significance tests of conditional mediated effects at sample statistics of the moderator(s), when sampling variabilities of such sample statistics are ignored versus accounted for. In the framework of a one-mediator model, design factors include magnitude of the a path (from the predictor to the mediator), magnitude of the b path (from the mediator to the outcome), strength of the moderation effect, sample size, and whether one or two mediational paths are moderated. Methods examined include joint significance test and the percentile bootstrap confidence interval approach.

(CAU) Causal Inference and Mediation

Poster 31: MEASURING THE HETEROGENEITY OF INDIVIDUAL-LEVEL TREATMENT EFFECTS WITH MULTILEVEL-OBSERVATIONAL DATA

Youmi Suk, University of Wisconsin, Madison; Jee-Seon Kim, University of Wisconsin, Madison

Multilevel latent class analysis and mixture propensity score models have been implemented to account for heterogeneous selection mechanisms and for making proper causal inference with observational multilevel data (Kim & Steiner, 2015). The scenarios imply the existence of multiple selection classes, and if class membership is unknown, homogeneous classes can be identified via multilevel logit latent class models. Multiple selection classes can be identified and units can be classified at the cluster level or at the individual level, depending on the nature of the heterogeneity. Using the Korea TIMSS 2015 data, this study examines potentially heterogeneous treatment effects of private science lessons by inspecting multiple selection classes (e.g., different motivations to receive the lessons) at the student level as well as at the school level. Procedures and implications for identifying selection and/or outcome classes at the individual level and cluster level are discussed.

(CAU) Causal Inference and Mediation

Poster 32: EVALUATION OF PERSON FIT IN COMPUTER ADAPTIVE TESTING (CAT)
Beyza Aksu Dunya, University of Illinois at Chicago

One of the most advantageous feature of CAT is targeting examinees' ability levels to maximize information and minimize error. One challenge associated with targeting is that traditional misfit statistics may not function effectively in CAT since it is unlikely that a person answers easier or harder item for his/her ability level. This simulation study elaborates on a specific source of misfit (lack of knowledge on a certain content by a sub-group of examinees) in a fixed-length CAT. 1PL IRT/Rasch dichotomous model was employed to calibrate difficulty and ability parameters for the simulated CAT (Rasch, 1960). Three methods for person misfit diagnosis were employed. As a traditional method, WINSTEPS mean-squared (MNSQ) person fit statistics was utilized to detect misfitting examinees. The second method, Wald-Wolfowitz runs test (Meijer & van Krimpen-Stoop) was based on comparing estimated and observed number of changes from correct to incorrect responses (or vice versa). In the last method, regression method, (Stahl; 2014), a regression line was plotted through the points representing a plot of ability estimate on the Y axis and sequence of items on the x-axis. The assumption is that the slope of the line should approach zero as the exam proceeds. Deviations from this slope would indicate examinees showing aberrant rise and fall in their estimated ability. My findings supported that it is even more challenging to detect misfit in CAT when only a subgroup of examinees are impacted by IPD and the problem of detecting person misfit in CAT still exists.

(CBT) Computer-Based Testing

Poster 33: AUTOMATED ITEM GENERATION USING MEDICAL DIAGNOSTIC INFORMATION
Ruben Castaneda, National Board of Osteopathic Medical Examiners; Qiuming Zhang, National Board of Osteopathic Medical Examiners

In all large-scale examinations, proper test construction depends largely on the quantity and statistical quality of the item pool. Automated test assembly requires a large item pool in order to have (a) a selection of items that meets the blueprint constraints and (b) items with acceptable difficulty parameters. Using automated item generation, we can maintain a large item pool while ensuring high quality. The research aims to automate the process of generating test items for the NBOME examination process. In this presentation, we first review the background and ideas of automated item generation, the procedure of collecting medical diagnostic information from websites, and the steps of generating medical content items for COMLEX-USA. To generate multiple-choice items, our method requires that we generate several templates of items with inputs for factors such as characteristics, demographics, and symptoms. We implement this procedure via R (for item generation) and Python (for web-scraping).Then, we discuss the use of the likelihood information collected from online medical websites to link the relationship between inputs and diagnoses. Using this likelihood information, we can automatically generate items with pseudo-difficulty parameters based on the relationship between the item key and distractors. This method also allows us to assign a value to the item based on the combinations of template inputs and their relationship to the key. The goal of this project is to create an alternative, efficient and sustainable method of generating items while simultaneously reducing costs associated with the process. Suggestions and next steps are discussed.

(CBT) Computer-Based Testing

Poster 34: SUBSCORE RELIABILITY CONSIDERATIONS IN MULTISTAGE ADAPTIVE TEST DESIGN
Yanming Jiang, Educational Testing Service

Background: Subscore reliabilities are generally low due to small numbers of items. To achieve reliabilities similar to linear tests, shorter subtests may be sufficient for a multistage adaptive test (MST) because of their adaptive nature. What subtest lengths would be adequate for K-12 large scale assessments with an MST design? Aim: We seek to determine the minimum subtest lengths required to meet target subscore reliabilities. It is expected that results of this study will help inform future MST design. Methods: We conduct a simulation study based on multidimensional IRT models. Augmented subscore reliabilities (Wainer et al., 2001; Edwards et al., 2006) will be estimated and compared with non-augmented reliabilities. Two-stage MSTs with one 1st stage and three 2nd stage modules are

considered. Simulation conditions:
• Test length: 40-56
• Number of subscores: 2, 3, and 4
• Subtest length: 12-24
• Subscore correlations: 0.7, 0.8, and 0.9
• Sample size: 5,000
A mixture of both dichotomously-scored and polytomously-scored items are considered. Test assembly of an MST is conducted using lpSolve software in R (Diao & van der Linden, 2011; Konis, 2009). Results: For an MST comprising three 14-point subtests with subscore correlations of 0.7, non-augmented and augmented subscore reliabilities were 0.79-0.80 and 0.84-0.85, respectively. Score augmentation provided more added-value when subscore correlations were high. When subtest lengths were greater than 20, subscore reliabilities were higher than 0.8 even without score augmentation. Conclusion: The number of subscores that an MST can support depends on test length, desired reliabilities, subscore correlations, item pool, and examinee ability distributions.

(CBT) Computer-Based Testing

Poster 35: A WEIGHTED ITEM SELECTION METHODS FOR LARGE-SCALE CAT
Gamze Kartal, University of Illinois at Urbana-Champaign; Tong Wu, University of Illinois at Urbana-Champaign; Hua Hua Chang, University of Illinois at Urbana-Champaign

The convenient and environmentally friendly computer makes large-scaled Computerized Adaptive Testing (CAT) a popular format in educational testing nowadays. Due to the challenges and difficulties of operating large-scaled CAT, many issues remain unsolved and the items selection strategies are always a popular research direction for CAT scientists. More and more researchers are conducting extensive research on item selection strategies based on ability level estimate and discrimination parameter to examine the most satisfactory item selection algorithm. A high rated debate topic of item selection is the importance of the order of estimation and the classification of. This can be easily achieved by modifying the sequence of and while doing simulation study. We propose a first and then selection approach and a mixture approach of and with weight constraint on both to detect which element is more important during the operation of large-scaled CAT. These approaches can improve the estimation of examinees' latent traits accuracy and the efficiency of CAT item pool usage. In addition, such approaches can give practitioners a good idea of what they should care more while they are choosing item selection algorithm in the operational CAT. The proposed approaches are evaluated through a simulation study for a large-scaled CAT and compare to constraint weighted a-stratification and Maximum Priority Index (MPI). Bias and Root-Mean-Square Error (RMSE) are the evaluate criteria.

(CBT) Computer-Based Testing

Poster 36: EXAMINING CONSTRUCT VALIDITY AND MEASUREMENT INVARIANCE IN COMPUTERIZED ADAPTIVE TESTS USING RASCH-SUBSCORES
Xinyu Ni, Teachers College Columbia University; Catherine Close, Renaissance Learning

Computerized adaptive tests (CAT) present factor analytic challenges due to their inherent nonrandom missingness in item responses resulting from students taking different items. As such, the purpose of this study was to examine the construct validity and measurement invariance across pre-kindergarten to grade 12 in a national reading CAT dataset- Renaissance's Star Reading. We conducted confirmatory factor analysis and multiple-group factor analyses (MGCFA) with SAS software using the domain scores estimated from the IPL theory, which solve the problematic missingness issue in CAT. The goodness-of-fit (GOF) indexes indicated that the unidimensional assumption in STAR Reading tests holds in each grade from grade K to grade 12 and the configural and metric invariances hold across grades implying the model construct and factor loadings are invariant across grades. The current study proposed a useful method of conducting confirmatory factor analysis using Rasch domain scores with an operational data from Star Reading. Difficulties of large-scale non-random missingness and the non-fixed set of domain items in CAT item responses can be overcome so as to provide much needed and

meaningful information to teachers and administrators in the decision-making process using computerized adaptive assessments results.

(CBT) Computer-Based Testing

Poster 37: DETECTION OF ITEM OVEREXPOSURE USING BAYESIAN CHANGE POINT ANALYSIS
Xiaodan Tang, University of Illinois at Chicago; Haiqin Chen, American Dental Association

Item overexposure in the computer adaptive testing (CAT) refers to the repeated exposure of popular items that might become compromised rapidly. It may result in a decrease in the items' actual difficulty, which would bias proficiency estimation (Georgiadou, 2007). The efforts to prevent the popular items from being overly exposed to examinees focus mainly on strategies adding a random component to the maximum information item selection method, and strategies based on assigning each item a parameter to control its maximum exposure. In the current study, we adopt the latter strategy and added an item drift indicator explained by a change point analysis (CPA) parameter to track item overexposure occurrence in a 3-PL IRT model. CPA is used to detect whether there are unexpected changes in longitudinal data. Zhang (2013) proposed a change point analysis to construct a realtime sequential item monitoring procedure to detect a time point when items were leaked. Different from his study, we conducted a simulation study to track the time when the drift happens and the magnitude of the change for both item difficulty and discrimination under the Bayesian framework. We applied a MCMC sampling approach to estimate the time parameter when item drift occurred by sampling from the conditional posterior distribution of a set of item parameters produced from the prior distribution using the log-likelihood function. The contributions of this simulation study are twofold: it reveals the time of item overexposure; it also provides information on whether item parameter drift is due to item overexposure.

(CBT) Computer-Based Testing

Poster 38: DIRECT EFFECTS IN STEPWISE LATENT CLASS ANALYSIS: IGNORE OR MODEL?
Jeroen Janssen, Leiden University; Saskia van Laar, CEMO, University of Oslo; Zsuzsa Bakk, Leiden University; Jouni Kuha, London School of Economics

Recently several stepwise approaches were proposed for latent class modelling with external variables, namely bias-adjusted three-step approaches (BCH, ML) and a more direct two-step approach as an alternative to the much debated one-step approach. Thus far very little is known about the performance of these approaches in the presence of direct effects of external variables on the indicators of latent class membership. While the three-step approaches cannot model this direct effect, the two-step approach can do so. The current study focuses on addressing two different research questions: on the one hand the effects in terms of parameter bias of not modeling direct effects are investigated for the different stepwise approaches, and on the other hand the performance of residual and fit statistics in identifying such effects are investigated for the approaches that are able to model them. To this purpose, an extensive simulation study is performed taking into account both one-step (gold standard), two-step and the (bias-adjusted) three-step models in various conditions containing one or multiple general or cluster-specific direct effects. The results show that the stepwise approaches are more robust against misspecifications than the one-step approach. Furthermore we report the power of various residual statistics to identify direct effects when present using the one and two-step approaches. The results show that when the sample size is large and measurement model is strong, these effects can be identified, in other conditions the power is not sufficient.

(CCC) Classification, Clustering and latent Class Analysis

Poster 39: EVALUATING METHODS FOR DETECTING MEASUREMENT NONINVARIANCE IN LATENT CLASS MODELS
Tessa Johnson, University of Maryland, College Park

Recent advances in the field of finite mixture modeling have allowed for the investigation of the effects of Measurement Noninvariance (MNI) and Differential Item Functioning (DIF) in Latent Class Models

(LCMs). Failing to account for DIF in LCMs leads to overextraction of classes and bias in structural parameters (Nylund-Gibson & Masyn, 2016; Masyn, 2017). However, testing for direct effects of covariates on measurement items in LCMs remains controversial. While model-based approaches appear superior to post-hoc testing (Cole, 2017), no formal evaluation of the various model-based techniques for detecting DIF in these types of models yet exists. The present study addresses this gap by comparing three prevailing methods: Multiple Group (MG-LCM; Collins & Lanza, 2010), stepwise Multiple Indicator Multiple Cause (MIMIC-LCM; Masyn, 2017), and a variation on stepwise MIMIC-LCM. A real data example, using bullying and peer victimization survey responses from the Health Behavior in School Aged Children study (HBSC; Iannotti, 2012), is presented. It is demonstrated that different MNI and DIF detection methods identify different direct effects, potentially altering the inference made regarding the effect of the covariate of interest on latent class membership. Guidelines for method selection and application of multiple comparison procedures are discussed.

(CCC) Classification, Clustering and latent Class Analysis

Poster 40: USING DECISION-THEORETIC MODELS TO LOCATE INDIVIDUALS WITHIN A HIGH-DIMENSIONAL CONSTRUCT

Michelle Lamar, Educational Testing Service; Malcolm Bauer, ETS

Process data from digital performance assessments may contain evidence of complex, often interacting multi-dimensional constructs. Mapping these data individually to construct dimensions can be prohibitively expensive while aggregate metrics frequently lose the signal in the noise. As an alternative approach we present theoretically-grounded cognitive models that predict decision making within the task given particular characteristics or levels of the construct dimensions. Individual performances can then be classified by comparative fit to the profile exemplar models. An application of this approach is presented for a game-based assessment of cross cultural competence (3C). The 3C construct encompasses how efficiently and accurately an individual learns to function in an unfamiliar social-cultural environment. 3C is thought to include a complex mix of aspects of cognition, metacognition, and personality. The instrument places players in a simulated culture which they must learn to navigate to achieve the game goals. The actions selected by players within the game and the resulting game states are modeled using a partially observable Markov decision process (POMDP). Multiple models are developed, representing a mix of high and low abilities on a collection of factors that relate to 3C. Classification is then accomplished using a maximum likelihood metric. To make the unequal-length play records comparable, a fit metric is developed based on the standardized log likelihood of actions taken as predicted by each model. Properties of the models, the classification procedures and results, and the fit metrics will be discussed.

(CCC) Classification, Clustering and latent Class Analysis

Poster 41: INTERPRETING THE DIMENSIONALITY OF SEVERAL TEST FORMS: STATISTICAL VERSUS SUBSTANTIVE

Alexandra Lay, The University of North Carolina at Greensboro; Terry Ackerman, University of Iowa

There are many different ways in which test dimensionality may be interpreted or defined. When regarding tests as multidimensional, meaning they measure multiple test taker attributes or skills, it is typically expected that there is a unique factor underlying each dimension which makes them distinct from one another. One definition of dimensionality, in which test dimensions are defined by expertly developed test specifications, may be referred to as substantive dimensionality (Walker, Azen, & Schmitt, 2006). Substantive dimensions may be defined by test specifications such as content areas, cognitive processes, depth of knowledge, or a combination of such specifications. Another form of dimensionality that may be considered is statistical dimensionality, which defines dimensions using statistical procedures such as factor analysis or cluster analysis, rather than expert insight. While a clear distinction can be made between these two types of dimensionality, it is reasonable to expect the substantive specifications to inform the statistical results (Zhang & Stout, 1991). These varying definitions of dimensionality raise the question: how does substantive dimensionality compare to the statistical dimensionality of real test data? This study serves as an exploration of that question through analysis of ACT science and math items from several test forms. Substantive dimensionality is defined using

specifications such as content, cognitive process, and depth of knowledge information, whereas statistical dimensionality is determined through the use of analyses in DIMPACK 1.0. Results of the study highlight the complicated relationship between statistical and substantive dimensions and implications for interpreting the observed scores.

(CCC) Classification, Clustering and latent Class Analysis

Poster 42: EXPLORATION OF LCA CONDITIONS IN LATENT CLASS WITH LOW PROPORTION
Sooyong Lee, University of Texas at Austin

The goal of this study is to determine the ideal research design for LCA when proportionally small latent classes are present. To achieve this goal, Monte Carlo simulation was conducted to examine which research design factors contribute the most to the accuracy of LCA models which contain latent classes consisting of only a small proportion of the total population. In order to provide specific research guidelines for practitioners, the interactions among the design factors were also examined. The simulations involved LCA models with four latent classes, with three research design variables considered: sample size, the number of indicators, and indicator quality. These factors were tested in terms of the size of the smallest latent class, which had four settings: 3%, 5%, 10%, and 20% of the population. It was found that, in general, a greater sample size and a larger number of high-quality indicators were beneficial to LCA models with proportionally small latent classes. When the smallest latent class was particularly rare (i.e., 3% or 5% of the population), stricter research conditions were required to yield less unbiased estimates. These findings can help researchers planning an LCA for uncommon or rare groups, especially when they cannot obtain a sample size large enough to ensure accurate model estimation (e.g., N=500 or more) because they can select other research design factors that can mitigate the negative effects of a small sample size and improve LCA performance.

(CCC) Classification, Clustering and latent Class Analysis

Poster 43: ACCOMMODATING LARGE NUMBER OF ATTRIBUTES FOR COGNITIVE DIAGNOSIS MODELS
Yan Sun, Rutgers University; Jimmy de la Torre, The University of Hong Kong

In many practical testing applications, the granularity of the attributes using cognitive diagnosis models (CDMs) determine the number of attributes to be measured. Specifically, the finer grained the attributes are, the more attributes are needed, and vice versa. In theory, the number of attributes estimated by a CDM is unlimited, but current constraints such as computer memory and existing algorithms practically limit the number of attributes that can be analyzed to not more than fifteen. The paper proposes the accordion approach (AP) as a solution to the problem in situations where it is reasonable to assume that the attributes can be partitioned into non-overlapping subsets. The AP focuses on diagnosing one subset of attributes each time while collapsing the remaining attributes into coarser attributes. Two possible situations where the AP can be applied to are as follows: 1) when the subsets of attributes are associated with different domains; and 2) when the diagnoses of a wide range of attributes happen at different time points. The impact of number of domains (2 or 3), number of attributes per domain (3 or 5), test length (60 or 120), item quality (low, medium, or high), and underlying CDMs (DINA or G-DINA) are examined in a simulation study. Results demonstrate that, compared to complete-profile estimation, the AP can achieve relatively high classification accuracy using more reasonable computation time.

(CCC) Classification, Clustering and latent Class Analysis

Poster 44: AN ATTRIBUTE-SPECIFIC ITEM DISCRIMINATION INDEX IN COGNITIVE DIAGNOSTIC ASSESSMENT
Wenyi Wang, Jiangxi Normal University; Lihong Song, Jiangxi Normal University; Shuliang Ding, Jiangxi Normal University

There are two basic set of item discrimination index to measure discriminatory power of an item (Rupp, Templin, & Henson, 2010). The first one is based on descriptive measures from classical test theory, such as the global item discrimination index, and the second index is based on information measures from item response theory, including cognitive diagnosis index (CDI; Henson & Douglas, 2005), attribute

discrimination index (ADI; Henson, Roussos, Douglas, & He, 2008), modified CDI and ADI (Kuo, Pai, & de la Torre, 2016). There lacks an item quality index as a measure of item's correct classification rates of attributes. The purpose of this study is to propose an attribute-specific item discrimination index as a measure of correct classification rate of attributes based on a Q-matrix, item parameters, and the distribution of attributes. First, the attribute-specific item discrimination index was introduced. Second, a heuristic method was presented using the new index for test construction. The first simulation results showed that the new index performed well in that their values matched closely with the simulated correct classification rates of attributes across different conditions. The second simulation study results showed that the heuristic method based on the sum of the attributes' indices yielded comparable performance to the famous CDI. The new index provides test developers with a useful tool to evaluate the quality of diagnostic items. It will be valuable to explore the applications and advantages of using the new index for developing item selection algorithm or termination rule in cognitive diagnostic computerized adaptive testing.

(CCC) Classification, Clustering and latent Class Analysis

Poster 45: ANALYZING EDITING BEHAVIORS IN WRITING USING KEYSTROKE LOGS

Mo Zhang, Educational Testing Service; Mengxiao Zhu, Educational Testing Service; Paul Deane, Educational Testing Service; Hongwen Guo, Educational Testing Service

In this study, we analyze students' editing behaviors extracted from keystroke logs in writing assessments. As suggested by the Hayes (2012) cognitive model, a writing process is a multidimensional construct including four connected dimensions - proposing, translating, transcribing, and reviewing and editing. Each dimension has distinct behavioral features. For example, idea generation and task preparation (i.e., proposing) generally associate with pauses at the start of writing and at sentence boundaries where writers stop and think about what to say. With the availability of keystroke logs, these theoretically-defined subconstructs can be estimated separately. We use data collected from a writing assessment on argumentation for analyses. Participants are mostly 7th to 9th graders in U.S. middle schools. First, for every log, each keystroke is automatically labeled to indicate an action type (e.g., insert, delete). Indications of editing behaviors (i.e., delete, backspace, cut, paste, jump, and replace) are further identified. Then, each log is evenly divided into n time units based on a fixed interval, and the editing action counts are calculated for each time unit. Autocorrelation of the count variables across time units will be tested. In addition to visual representations, linear regression model will be fitted to the count variables. The extent and trend of the editing behaviors for a writer can then be respectively estimated by the intercept and slope of the model. Results of this study will have implications for classifying students' editing styles.

(CCC) Classification, Clustering and latent Class Analysis

Poster 46: SIMULATION STUDY OF SCORING METHODS FOR VARIOUS MULTIPLE-MULTIPLE-CHOICE ITEMS

Sayaka Arai, The National Center for University Entrance Examinations; Hisao Miyano, The National Center for University Entrance Examinations

The multiple-choice (MC) format is the most widely used format in objective testing. The "select all the choices that are true" item, which also called as multiple-multiple-choice (MMC) item, is a variation of the MC format, which has no instruction to indicate the number of correct choices. Although many studies have developed and compared scoring methods for the MMC item, the results have often been inconsistent. Arai & Miyano (2017) proposed new scoring methods and compared the scoring features of them. They also conducted numerical simulation, however they showed only one example, which has five choices with three correct choices. In this study, we conducted numerical simulations of other patterns of MMC items and examined the relationships between examinees' abilities (true score) and scores given by scoring methods. We illustrated the influence of the numbers of total choices and correct choices on each scoring method.

(CTT) Classical Test Theory

Poster 48: EFFECT OF MISSING DATA ON DETECTING NONUNIFORM DIFFERENTIAL ITEM FUNCTIONING

Onur Demirkaya, University of Illinois at Urbana-Champaign; Ummugul Bezirhan, Teachers College, Columbia University; Hua-Hua Chang, University of Illinois at Urbana-Champaign

Missing data and differential item functioning (DIF) are two commonly encountered situations in educational assessment. While implementations of DIF procedures require complete data matrices, frequently used approaches for treating missing data, zero imputation (ZI) and listwise deletion (LD), provide biased estimates unless the missing data are completely random. The research addressing this problem in the literature have heavily focused on examining the impact of proposed missing data techniques on uniform DIF while relatively little work has examined this issue in the context of nonuniform DIF detection. The goal of this simulation study is to extend the earlier literature by focusing on impacts of particular missing data techniques (listwise deletion, zero imputation, two-way imputation and response function imputation) on commonly used nonuniform DIF detection procedures (logistic regression, item response theory likelihood ratio test, and crossing SIBTEST) under three missing data mechanisms (MCAR, MAR, MNAR). The two-parameter logistic model will be used to generate data under three manipulated factors, including sample size ratio, missing data mechanisms, and percent of examinees missing data. Type I error and DIF detection rates will be compared to the baseline condition to determine optimal method under given conditions. The results of this study will assist practitioners and researchers in making an informed decision when using the missing data techniques in their DIF analysis.

(DIF) Measurement Invariance and DIF

Poster 49: NEW EFFECT SIZE MEASURE TO DETECT MEASUREMENT NON-INVARIANCE

Heather Gunn, Arizona State University; Kevin Grimm, Arizona State University

Introduction: Measurement invariance is a property that exists when a scale functions equivalently across groups (or across time in a longitudinal context). Without measurement invariance, valid group comparisons are not possible. Invariance is examined by assessing the fit of a sequential series of models; however, the statistical significance of differences in model fit is influenced by many factors, including sample size. Additionally, statistical significance is not always a strong indicator of the practical significance of the group differences. Effect sizes, on the other hand, are independent of sample size and can be used to determine the magnitude and practical importance of any lack of measurement invariance. Many effect sizes of measurement non-invariance have been developed in item response theory (IRT), but there are only a few effect size measures of measurement non-invariance in the factor analytic (FA) framework. Method: We developed an effect size measure of measurement non-invariance in the FA framework. A simulation study was conducted to evaluate the performance of the new effect size measure. We varied characteristics that influence invariance tests: sample size of the groups, test length, number of non-invariant items, direction of non-invariance, size of non-invariance, and type of non-invariance (i.e., metric invariance, scalar invariance). Results: We investigated the accuracy and behavior of the new effect size measure to detect non-invariance.

(DIF) Measurement Invariance and DIF

Poster 50: DIFFERENTIAL ITEM FUNCTIONING IN THE DEPRESSION SCALES BY GENDER

Jihye Kim, Kennesaw State University

Validity and reliability are the most fundamental elements of instrument justification. In addition, it is important to establish the degree to which the instrument performs uniformly regardless of specific groups' characteristics such as gender, race/ethnicity, etc. Differences in instrument performance across groups can lead to errors in screening individuals, inaccurate estimates in research, and lack of fairness and equity between groups. Thus, it is vital to assess if the instrument performs consistently across groups, a property referred to as differential item functioning (DIF). This research focuses on examining the Depression Scale of the Profile of Mood Status Short Form (POMS-SF) and the Center for Epidemiologic Studies - Depression (CES-D) to determine whether these instruments measure depression symptom consistently and accurately particularly by gender. One reason this is of interest is

that numerous studies have found higher levels of depressive symptoms among women than among men. The purpose of the study is to determine whether the depression Scales of the POMS-SF and the CES-D can identify the presence of depression accurately and differences in levels of depression by gender. The results will contribute to mental health research as well as demonstrate the potential of archival research in survey item analyses. By improving the accuracy rates of existing instruments, researchers might be able to develop better theoretical models in screening individuals for symptoms of depression.

(DIF) Measurement Invariance and DIF

Poster 51: ASSESSING MEASUREMENT INVARIANCE ON ZERO-INFLATED MEASURES
Mirim Kim, Texas A&M University; Myeongsun Yoon, Texas A&M University

To test measurement invariance (MI), factorial invariance (Meredith, 1993) is assessed through multiple-group confirmatory factor analysis (MGCFA) to see whether the targeted latent trait is measured without a measurement bias across subgroups (Vandenberg & Lance, 2000). Compared to lots of studies on MI, not many studies on how non-normality affects detecting MI have been studies, therefore, assessing factorial invariance with non-normal data might be difficult for researchers to approach. This study tested MI of 10 bullying/ victimization items of Korean Youth Panel Survey in 2003. Lots of students answered 'no experience', which means 'zero' response, therefore, the data was extremely inflated to zero, and other responses were skewed to small number of experiences. MGCFA and two-part factor model (Kim & Muthen, 2009) were applied, and their results were compared. Although different approaches to select a reference variable were used for two models (Jung & Yoon, 2017; E. S. Kim, Joo, Lee, Wang, & Stark, 2016; Rivas, Stark, & Chernyshenko, 2009), the forward approach with simultaneous comparisons were considered to see MI (Cheung & Lau, 2011; Jung & Yoon, 2016). Two models had different results. Two-part factor model showed that scalar invariance held across groups, but MGCFA showed the partial invariance. This study not only demonstrated a practical application of two-part factor model to detect MI when it comes to zero-inflated measures, but also showed problematic issues from non-normality with respect to MI testing.

(DIF) Measurement Invariance and DIF

Poster 52: MEASUREMENT INVARIANCE IN WITHIN-SUBJECTS DESIGNS
Saskia van Laar, CEMO, University of Oslo; Johan Braeken, CEMO, University of Oslo

Measurement invariance implies that as long as the same measurement instrument is being used, scores can be naturally compared. For between-subjects designs, procedures and tests for detecting differential item functioning (i.e., non-invariance) between groups of different subjects are relatively well-established. Although the terminology is largely domain-specific, the underlying principle is the same: A stepwise procedure is followed that investigates the viability of restricting specific measurement model parameters to be equal across groups. For within-subjects designs these procedures and tests have been less explored. Here, it is about fair comparisons between the same group of subjects across repeated measurements. Modifications to the default measurement invariance procedures might be required to account for this difference in design. What with the theoretically expected local dependence between repeated measurements of the same item? Should the independence null reference model for model comparisons and fit indices be changed to a longitudinal variant, and if so to which one? We will present results on a simulation study investigating and contrasting the performance to detect differential item functioning in a within-subjects design of (i) a default between-subjects measurement invariance approach and (ii) 2 variants of proper within-subjects measurement invariance approaches.

(DIF) Measurement Invariance and DIF

Poster 53: EXPLORATORY BIFACTOR ANALYSIS WITH SMALL SAMPLE SIZES
Dong Gi Seo, Hallym University; Sunho Jung, Kung-Hee University

The application of exploratory factor analysis (EFA) is common in the behavioral sciences. With small sample sizes (roughly 50 or fewer observations), two approaches have been employed for EFA: factor analysis using unweighted least squares (ULS-FA) and regularized exploratory factor analysis (REFA). Exploratory bifactor analysis has recently gained popularity as a technique for identifying distinct domain-specific factors uncorrelated with a broad common factor. The Schmid-Leiman orthogonalization procedure has proven to be a popular tool for obtaining a bifactor solution. However, there is no evidence regarding which approach should be preferable for exploratory bifactor modeling in small-sample situations. The authors conduct a comprehensive simulation study to evaluate the small sample behavior of the two approaches in terms of bifactor structure recovery under different experimental conditions of sample size, factor loading, number of variables per factor, number of factors, and factor correlations. The results show that REFA is recommended for use over ULS-FA, particularly under difficult conditions of low factor loadings and/or few factors with a small number of measured variables.

(ECM) Estimation and Computational Methods

Poster 54: ASSESSING MEASUREMENT INVARIANCE WITH TESTLET EFFECT UNDER MCCFA FRAMEWORK
Xiangyi Fu, Sun Yat-sen University

Testlet format tests are broadly used in educational tests like reading and comprehension assessments. This article investigated testlet effect under the multiple-group categorical CFA (MCCFA) framework. Specifically, we investigated the influence of sample size, magnitude of variance in threshold, magnitude of loading, impact of latent factor, and different type of testlet effect on rejection rate (Type I error or power) using DIFFETST, Wald test, and changes in approximate fit indexes (AFI). The results of Type I error showed all methods were acceptable under different conditions. However, the result of power showed that changes in AFI were not good in detecting variance. Wald test was better in detecting measurement variance. Moreover, larger sample size, larger difference in thresholds, small loading, and no impact in latent factor would have larger power.

(FAC) Factor Analysis

Poster 55: ON THE EXTENDED GUTTMAN CONDITION
Kentaro Hayashi, University of Hawaii at Manoa; Ke-Hai Yuan, University of Notre Dame; Ge Jiang, University of Notre Dame

It is well-known that factor analysis and principal component analysis often yield similar estimated loading matrices. Guttman (1956) identified a condition under which the two matrices are close to each other at the population level. We discuss the matrix version of the Guttman condition for closeness between the two methods. It can be considered as an extension of the original Guttman condition in the sense that the matrix version involves not only the diagonal elements but also the off-diagonal elements of the inverse matrices of variance-covariances and unique variances. We also discuss some implications of the extended Guttman condition.

(FAC) Factor Analysis

Poster 56: EFFECT OF PROCESSING SPEED AND ITS REPRESENTATION ON MODEL FIT
Karl Schweizer, Goethe University Frankfurt

Processing speed is a source of performance that is mostly neglected in psychological assessment although in the case of a time limit for testing processing speed is likely to influence the outcome of assessment (Oshima, 1994) and to impair its validity (Lu & Sireci, 2007). The reason is that a time limit together with lack of processing speed can prevent a number of participants from reaching their highest possible score. In a simulation study strong, medium and weak effects due to a time limit on

model fit were investigated in sets of 500 matrices of structured random data respectively. The effects were simulated by means of the logistic function as percentages of omissions. It turned out that impairment was signified by some fit indices whereas other fit indices always indicated good model fit. Furthermore, the possibility to improve model fit by means of two-factor models of measurement with free and fixed factor loadings was investigated. Fixed factor loadings represented the effect in a way that reflected how it unfolded in the sequence of the items of a scale. Although the results of this investigation indicated a large improvement in model fit, the degree of good model fit for data with no effect was not reached. The results for the two models did not differ from each other.

(FAC) Factor Analysis

Poster 57: BIFACTOR STRUCTURE OF CHILDREN'S COPING FOR PARENTAL DIVORCE OR DEATH

Jenn-Yun Tein, Arizona State University; Yu Liu, University of Houston

The current study examines a bifactor model for coping strategies among children who experienced parental death and children who experienced parental divorce. Children may apply active coping strategies (e.g., seeking understanding, positive cognitive restructuring) or avoidant coping strategies (e.g., repression, wishful thinking) when facing stressful events. Active coping strategies are theorized as adaptive coping strategies that are hypothesized to alleviate negative mental health outcomes, whereas avoidant coping strategies are theorized as maladaptive coping strategies that are hypothesized to aggravate negative mental health outcomes. However, these two types of coping strategies are highly positively correlated, which may imply the existence of a general factor across all coping strategies. The aim of this study is two-fold. First, we will examine the latent structure of coping strategies with a bi-factor modeling approach that separates a general coping factor from the unique factors of active coping and avoidant coping. Second, we will examine measurement invariance of the bi-factor structure between two samples -- children who experienced parental death and those who experienced parental divorce. Parental death and parental divorce are the two most stressful events for children, and it is unknown whether the coping structures are similar for these two groups. This study uses data on 240 children from the New Beginning Project (for children of divorce; Wolchik et al., 2000) and 244 children from the Family Bereavement Project (for children who face parental death; Sandler et al., 2003) to examine the aim.

(FAC) Factor Analysis

Poster 58: A BI-FACTOR MODEL OF INSTITUTIONAL INTEGRATION SCALE

Xiaoyan Xia, University of Pittsburgh

We evaluate the factor structure of the Institutional Integration Scale (IIS; Pascarella & Terenzini, 1980) by examining the following models: three bifactor models with one general integration factor and specific factors representing 1) five specific integration dimensions, 2) two specific factors representing social and academic integration, 3) two specific factors representing faculty and student integration. Additionally, three corresponding first-order CFA models were examined including five-factor CFA representing the originally proposed multidimensional factorial structure, and two two-factor CFA, one measuring social and academic integration (Tinto's model), and the other faculty and student integration (French and Oakes' model). We also assess the measurement invariance by gender and race, as well as the predictive validity of integration. The implications of these findings for interpretation and use of this scale are discussed.

(FAC) Factor Analysis

Poster 59: COMPARISON OF GROUP-BASED AND MODEL-BASED APPROACHES TO PERSON-FIT ANALYSIS

Yu Bai, Teachers College, Columbia University; Young-Sun Lee, Teachers College, Columbia University

Person-fit analysis is widely adopted to detect the discrepancy between an examinee's test score and his/her true ability. Two approaches can be distinguished in person-fit research: group-based approach

and model-based approach. In the group-based approach, an item score pattern is evaluated given the item score pattern of other people in the group or given that a nonparametric model fits the data. In the model-based approach, the likelihood of an item score pattern is evaluated given a parametric model fits the data. While under CDMs, the common practice of person-fit evaluation relies on the model-based approach, results from previous studies also suggested the superiority of group-based indices in identifying aberrant-responding examinees (Karabatsos, 2003). In this study, the person-fit statistics were compared under the DINA model in terms of the performance of detecting spurious high scores and spurious low scores. 40 datasets were simulated according to a fully-crossed 5 by 2 by 2 by 2 design: 5 intensities of aberrant responding examinees (0% of aberrant examinees, 5% of aberrant examinees, 10% of aberrant examinees, 30% of aberrant examines, and 50% of aberrant examines); 2 types of aberrant behaviors (spuriously high, A=1; spuriously low, A=0), 2 types of item quality (high quality, s=g=0.1; low quality, s=g=0.3), and 2 test length (short, J=30; long, J=60). Each dataset consisted of 100 simulated examinees and 100 replications were done. Results indicated that when the models were correctly specified, group-based person-fit index, ZU3, performed less optimal than model-based person-fit indices, joint LRT and marginal LRT.

(FCM) Model Fit, Comparison and Diagnostics

Poster 60: AN INTEGRATIVE FRAMEWORK OF MODEL EVALUATION
Wes Bonifay, University of Missouri; Li Cai, University of California, Los Angeles

The present work considers three schools of thought - frequentist statistics, Bayesian inference, and information theory - that seem to offer philosophically and methodologically dissimilar perspectives on statistical model evaluation. Though these schools may seems at odds, the present work develops a simple theoretical framework that integrates each perspective. Central to this unified framework is the realization that these perspectives differ along two dimensions. The first dimension reflects the treatment of the data in model evaluation: frequentist methods typically rely on the observed data, Bayesian methods involve replicated data, and information-theoretic methods consider all possible data. The second dimension reflects whether or not the method of evaluation involves re-fitting the model (e.g., the parametric bootstrap technique involves reliance on the observed data and re-fitting the model, while predictive model checking procedures use replicated data and do not require re-fitting the data). When arranged along these two dimensions, these seemingly disparate approaches to model evaluation are not so dissimilar. An extended example demonstrates how to use this framework to evaluate—from frequentist, Bayesian, and information-theoretic perspectives—whether a specific item response theory model is appropriate for an empirical data set. Conclusions drawn from any one perspective would be limited in scope, but an evaluation based on all three perspectives is comprehensive, enlightening, and difficult to refute. Ultimately, by identifying common ground between the frequentist, Bayesian, and information-theoretic methodologies, this integrative framework will lead to a better understanding of psychometric model evaluation.

(FCM) Model Fit, Comparison and Diagnostics

Poster 61: EXAMINING PERFORMANCE OF THE WRMR WITH CATEGORICAL AND CONTINUOUS DATA
Ning Jiang, University of South Carolina; Jin Liu, University of South Carolina; Dexin Shi, University of South Carolina

In Structural Equation Modeling (SEM) techniques, fit indices are used to describe the fit of a model and also to provide support for other decisions. The Weighted Root Mean Square Residual (WRMR), which is a relatively new fit statistic, is increasingly used for support of model-data fit. However, not much is known about the performance of this index. This study aims to fill this gap in the research and will use Monte Carlo method to examine the performance of the WRMR when categorical and continuous data are analyzed. Other conditions (model size, sample size, magnitude of factor loadings, item distribution, and misspecification) will also be simulated to investigate the sensitivity of the WRMR. Mplus (v. 8) will be used with the robust estimators of WLSMV (Mean- and Variance-adjusted Weighted Least Square) and MLMV (Mean- and Variance-Adjusted Maximum Likelihood) corrections. One thousand replications will be run for each design cell, and replications which show convergence issues or improper solutions will be removed for further analysis. In addition, the WRMR will be compared relative to other well-known fit indices (i.e., global $\chi 2$, SRMR, TLI, CFI, and RMSEA). This study can provide applied researchers

information about situations where the index works appropriately and under which conditions they should be wary about using WRMR.

(FCM) Model Fit, Comparison and Diagnostics

Poster 62: DETECTING ITEM PREKNOWLEDGE WITH A SMALL NUMBER OF COMPROMISED ITEMS
Hwanggyu Lim, University of Massachusetts Amherst; Scott Monroe, University of Massachusetts Amherst

Sinharay (2017) suggested a signed likelihood ratio test statistic, L_s, and a signed score test statistic, R_s, for detecting aberrant response patterns when a set of compromised items can be identified. In a simulation study using a nonadaptive assessment, the number of compromised items in the set, n_c, was at least 10. In practice, however, n_c may be smaller. But, because L_s and R_s have an asymptotic standard normal distribution under the null hypothesis (H0), small n_c may result in poor calibration. One solution for this problem is to use a simulation-based empirical null distribution by adapting the generalized resampling-based approach (Sinharay, 2016). Therefore, this study investigated the performance of L_s and R_s for detecting aberrant examinees for small n_c, using both the frequentist approach with the normal assumption under H0, and the adapted resampling-based approach. In the simulation study, we manipulated test length and size of n_c (3 and 5). We employed the 3PL IRT model and estimated abilities using the WLE. For resampling, a MCMC method using the Metropolis-Hastings algorithm was used. In the results, Type I error rates of both statistics with the resampling-based approach were close to nominal levels of .05 and .01, while those of L_s with the frequentist approach were inflated at the nominal level .05 for n_c=3. Power of both statistics with the resampling-based approach were adequate. Thus, the results support that the use of the resampling-based approach for detecting item preknowledge is promising for small n_c.

(FCM) Model Fit, Comparison and Diagnostics

Poster 63: EXAMINING BAYESIAN APPROACHES FOR INVESTIGATING MEASUREMENT INVARIANCE ACROSS TWO GROUPS
Yuanfang Liu, University of Cincinnati; Mark H. C. Lai, University of Cincinnati

Bayesian methods are becoming popular in psychological and educational research. However, there is limited studies evaluating the performance of Bayesian methods in detecting violations of measurement invariance, which is a prerequisite for valid cross-group comparisons on means and associations of constructs. Aside from the use of vague priors that generally yields similar parameter estimates as with the frequentist approach, Muthén and Asparouhov (2012) suggested to place a small variance prior (e.g., normal with a mean 0 and a variance 0.01) on the difference of factor loadings (or intercepts) across groups to reflect a strong prior belief of equal loadings (or intercepts). In this study we systematically examined how Bayesian model evaluation indices, namely the deviance information criterion (DIC) and the posterior predictive p value, perform in detecting violations of measurement invariance, compared to frequentist test statistics and fit indices. A simulation was conducted with varying sample size, magnitude of non-invariance, and various choices of prior distributions in evaluating metric and scalar invariance for a one-factor model with two groups. Preliminary results showed that DIC provided better results than frequentist fit indices in balancing high detection rate and low false positive rate of measurement invariance, especially when sample size is large. The use of small variance priors resulted in higher detection rates of DIC with some added benefits. Practical recommendations for evaluating measurement invariance under the Bayesian framework will be discussed.

(FCM) Model Fit, Comparison and Diagnostics

Poster 64: ANALYSIS OF GRAPHICAL DIAGNOSTIC CLASSIFICATION MODEL WITH COVARIATES
Ummugul Bezirhan, Teachers College, Columbia University; Young-Sun Lee, Teachers College, Columbia University

A graphical diagnostic classification model (GDCM) is proposed by Kang, Liu, and Ying (2017) to account for local item dependency by incorporating a Markov network to diagnostic classification models

(DCMs). DCMs are designed to obtain information about examinee's mastery profiles on fine-grained skills. Models that include covariates provide additional information about factors that can explain examinee's skill profiles and their performances. Therefore, this study aims to extend the GDCM framework by including covariates to the model at both item level and attribute level. A simulation study is conducted in order to illustrate the performance of the graphical model with the presence of covariates and compare the results against those obtained from standard DCM. DINA model is implemented to simulate response data on varying graphical structure, distribution of covariates and sample size. The test length and attribute numbers are the fixed factors. Root mean square deviation (RMSD) and bias are used to assess the quality of the parameter recovery. The result of this study will assist researchers in the use of covariates to gain supplementary diagnostic information related the factors that affect items and attributes.

(GRM) Graphical Model

Poster 65: ESTIMATING PSYCHOLOGICAL NETWORKS WITH THE BAYESIAN BOOTSTRAP
Donald Williams, University of California, Davis

An important goal for psychological science is developing methods to characterize relationships between variables. A common approach uses structural equation models to connect latent factors on a structural level to a number of observed measurements. More recently, network models have been developed that provide an alternative approach for characterizing conditional relationships among variables with the precision matrix. Whereas classical (i.e., frequentist) methods such as GLASSO are commonly used in psychology, Bayesian methods remain relatively uncommon in practice and methodological literatures. Here we propose a Bayesian method that uses probability weights drawn from a Dirichlet distribution over the input data to estimate the precision matrix. These probability weights provide a bootstrap estimator for the posterior, which is computationally cheap compared to Markov chain Monte Carlo sampling. Specifically, we introduce two models based on a non-regularized and regularized approach, both of which rely on directional posterior probabilities for determining non-zero relationships. The latter allows for estimating network models when the number of variables (p) exceeds the number of observations (n). With numerical experiments, we demonstrate that performance often exceeds that of classical methods with respect to correctly identifying conditional relationships. In addition, both models show similar results for frequentist risk measured with Kullback–Leibler divergence. We discuss implications for psychology, as well as the Bayesian literature on the topic of estimation in high-dimensional settings.

Poster 66: TOWARDS THE IMPORTANCE OF RULE-BASED ITEM CONSTRUCTION USING LLTM-R
Tobias Alfers, University of Vienna, Austria; Georg Gittler, University of Vienna, Austria

Since the beginning of adaptive testing researchers strive to establish a procedure for item selection from a hypothetically infinite item universe. The importance of new approaches like Automatic Item Generation (AIG) for the development and validation of such an item universe have already found its way into the minds of many psychometricians. Strong drivers for the research of developing item generators are the promised guarantee for maintaining test security and the cost-effective development of new items. There is a recognizable trend to build such item generators before item contents have been sufficiently validated, which is a serious threat to the quality of newly developed computer-based psychological tests. This manifests in a poor understanding of the determinants of item difficulty and ultimately in puzzling why some items do not perform as intended. In this research we present the construction of the 3DW, a three-dimensional cubes test consistent to the Rasch model, as a best practice example for applying strict rule-based item construction procedures. We utilized and promote the linear logistic test model and its extensions within the Bayesian framework to evaluate the model fit at both test- and item-level (as stated by Janssen et al., 2004), which allows us to strictly focus on the content validity of the test instrument. This research clearly points out that it should be the primary goal and a necessary condition to extensively and thoroughly test the psychometric qualities of an item-type, before moving further to build an item generator.

(IRT) Item Response Theory

Poster 68: HOW DOES GUESSING INFLUENCE THE ABILITY ESTIMATION
Yanhong Bian, Rutgers, the State University of New Jersey; Werner Wothke, American Councils for International Education

Guessing behavior has been a widely acknowledged problem in multiple-choice tests especially for test takers with relatively lower abilities (Andrich, Mariais & Humphry, 2012; Roediger & Marsh, 2005). Even though the guessing problem in multiple-choice tests has been proposed since the 1970s and 1980s (Harden, Brown, Biran, Ross & Wakeford, 1976; Hambleton, 1982), the influence of the guessing parameter on the ability estimation has never been deeply explored. An interesting problem was found in the actual educational testing that using the response datasets for two groups taking the same test, the calibrated guessing parameters can differ a lot, with one showing little or no guessing and another one showing high guessing. What will happen to the ability estimates if the wrong item response theory (IRT) model is applied to the response data? In this study, the impact of model selection in regards to guessing on the ability estimation under IRT framework was investigated using both simulation study and an analysis of actual educational testing data. Two models were utilized for the ability estimations, the two-parameter logistic (2PL) model and the three-parameter logistic (3PL) model. Different test length, number of examinees, ability distributions and item guessing levels were considered in the simulation study. The results show that the ability estimates using models with and without guessing parameters have a linear relationship, especially when the true guessing parameters are small.

(IRT) Item Response Theory

Poster 70: TO IRTREE OR NOT TO IRTREE: INVESTIGATING TREE-STRUCTURE RECOVERY
Dries Debeer, University of Zurich

In recent years, item response tree models (also known as IRTrees; De Boeck & Partchev, 2012; Böckenholt, 2012) have become a popular modeling and research tool in education measurement. For instance, they have been utilized for modeling response styles, slow/fast intelligence, missing/omitted responses, answer change behavior, etc. IRTrees model categorical outcomes using a tree structure of sequentially interconnected subprocces. Implicitly they represent a belief about the process underlying the outcome/item response. Commonly, only one IRTree is applied, and its structure is based on substantive knowledge (i.e., on a belief about the response process). However, recently several authors suggested that fit indicators, such as the AIC, can be used to compare multiple IRTrees with the same number of subprocesses/parameters, in order to decide which tree structure best describes the response process (Böckenholt, 2017; Debeer, Janssen & De Boeck, 2017). Polytomous item responses can be generated according to equally-sized but differently structured IRTrees. Yet, can model fit measures indeed be used to identify the true data generating IRTree structure? Or do IRTrees (with an equal number of subprocesses/parameters) provide equal fit to the data? Using an extensive simulation study, we critically investigate whether, and under which conditions, the IRTree approach can recover the data generating tree structure. In addition, the potential of Vuong's test (1989) for comparing the fit of non-nested IRTrees is discussed.

(IRT) Item Response Theory

Poster 71: COMPARISON OF THREE METHODS TO ASSESS ADEQUACY OF PARAMETRIC IRF
John Donoghue, Educational Testing Service; Adrienne Sgammato, Educational Testing Service

Item response theory enables several important applications, such as IRT-based linking/equating and CAT. The accuracy of these procedures is contingent on the model assumptions being satisfied. One key assumption is the form of the item response function (IRF). This study compares the accuracy and power of three popular item fit measures:
• Douglas and Cohen compare the parametric IRF to a non-parametric computed as root integrated squared error (RISE) between the curves. Significance is based on parametric bootstrap.
• Orlando and Thissen's S-X2 compares observed and expected counts for each total score, resulting in a Pearson chi-square, which is compared to a chi-squared distribution.
• Stone's X^2* uses the "pseudo-counts" that arise during EM estimation to compute observed and expected values for a Pearson or likelihood ratio chi-squared statistic, which is approximated as scaled

chi-square. Parametric simulation estimates the degrees of freedom and scaling factor.
This simulation compares these three methods for dichotomous IRT. Items that fit the model are generated using 3PL. Misfit is based on cubic splines fit to 8 empirical items.
Factors manipulated:
• Test length 5, 10, 15, 20, 30, 40, 60 and 80 items
• Percent of items demonstrating misfit, 0%, 10% or 20%
• Sample size: 500, 1000, 2000, 3500, and 5000.
• Significance test critical value: determined by parametric bootstrap (all measures) or nominal distribution (only S-X2 and X^2*).
• "Tuning parameters" of the procedures (e.g., bandwidth, minimum expected frequency) are also manipulated.
The results of the study will inform decisions about how best to assess IRF fit in applied settings.

(IRT) Item Response Theory

Poster 72: A COMPARISON OF IRT SCALE TRANSFORMATION RESULTS UNDER VARIOUS CONDITIONS

Youngjin Han, Yonsei University; Juyoung Jung, Yonsei University; Guemin Lee, Yonsei University

It has been widely believed that anchor test need to be a representative of the total test in respect of both the content and the statistical characteristics when different groups of examinees administering different test forms are to be equated (Kolen & Brennan, 2014). Recently, it has been reported that anchor tests with narrower distribution of item difficulty compared to the total test tend to function more accurately with respect to equating bias and standard error (Liu et al., 2011a, 2011b; Sinharay & Holland, 2006, 2007). In this study, the consistency of this tendency that the anchor tests with relaxed statistical assumptions function appropriately was investigated by comparing results of IRT scale transformation under various simulation conditions. It is expected to give operational convenience in various testing situations allowing more flexible standards in constructing anchor test if the consistency holds. The study considered 3 types of anchor test including mini test, semi-midi test, and midi test according to spread of item difficulties. 3 conditions regarding mean difficulty of anchor test and non-equivalency of examinee ability were considered respectively. In all, 27 conditions were examined. In order to minimize the influence of the scale transformation method, the scale transformation was performed by the Stocking-Lord method (Stocking & Lord, 1983) after separate calibration and an item characteristic curve criterion (Hanson & Beguin, 2002) was used to evaluate the accuracy on results of scale transformation.

(IRT) Item Response Theory

Poster 73: COMPARATIVE EVALUATION OF THE GRADED RESPONSE AND FACTOR ANALYSIS MODELS

Takahisa Ikeda, Senshu University; Kensuke Okada, Senshu University

In psychology, many studies that measure personality traits collect responses through psychological questionnaires. Typical applied studies conduct a factor analysis on the basis of the observed item response data for measuring the common latent factor. However, the ordinary factor analysis model assumes that the observed variables are continuous and realized from the underlying normal distribution. Thus, the factor analysis model may not appropriately represent the data generating mechanism of the observed Likert scale questionnaire item responses. On the other hand, the graded response model represents the underlying data generating mechanism of the ordinal categorical item responses. Thus, theoretically, this model would be more suitable for psychological questionnaire data analysis. In view of the aforementioned theoretical consideration, the objective of the current study is to quantitatively evaluate the difference in performance between these two models. For this purpose, we conducted a simulation study. Specifically, we generated artificial questionnaire data, fitted them to these two models, and compared their performance in parameter recovery and prediction. The results suggest the potentially large-scale applied utility of the graded response model.

(IRT) Item Response Theory

Poster 74: AN ITEM RESPONSE MODEL FOR DISCRETE OPTION MULTIPLE CHOICE ITEMS
Nana Kim, University of Wisconsin, Madison; Daniel M. Bolt, University of Wisconsin, Madison; James Wollack, University of Wisconsin, Madison; Yiqin Pan, University of Wisconsin, Madison; Carol Eckerly, Alpine Testing Solutions; John Sowles, Ericsson, Inc.

A new format for computer-based administration of multiple-choice items, the discrete option multiple choice (DOMC) format, is receiving growing attention due to potential advantages related both to item security and control of test-wiseness. A unique feature of the DOMC format is the potential for an examinee to respond incorrectly to an item for different reasons -- either failure to select a correct response, or incorrect selection of a distractor response. The feature motivates consideration of a new item response model that introduces an individual differences trait related to general proclivity to select response options. Using empirical data from an actual DOMC test, we validate the model by demonstrating the statistical presence of such a trait, and discuss its implications for test equity and the potential value (need) for added item administration constraints.

(IRT) Item Response Theory

Poster 75: ITEM WORDING DIRECTIONALITY EFFECTS ON PANAS SCALE USING POLYTOMOUS IRT
Sohee Kim, Oklahoma State University

Psychological scales such as depression, anxiety, and stress inventories tend to consist of positively- and negatively-worded items with a Likert-type response format. For these items, the polytomous item response theory (IRT) methods most often used are the graded response (GR) and generalized partial credit (GPC) models. More recently, but less commonly, the nominal response (NR) model is being applied to truly ordinal response data such as a Likert-type response (Preston, Reise, Cai, and Hays, 2011). Since the assumption of a common discrimination across categories by the GR and GPC models may be violated for many items on a scale. The NR model allows for the investigation of how adjacent categories may discriminate differently when items are positively or negatively worded and is useful to better understand distinctions between response categories on self-reported psychological scales (Preston et al., 2011). Thus, this study investigates the effects of item wording directionality by using different polytomous IRT methods: GR, GPC, and NR models. For the study, the response data were collected from 878 participants completing the 20-item PANAS scale. This scale was divided into two parts: 10 items measuring Positive Affect and 10 items measuring Negative Affect. The items on this scale consist of a list of words that describe different feelings and emotions. In order to evaluate the effects of item wording directionality, model fit, test information curve, estimated parameters, item information curve, and category characteristic curve are compared for three polytomous IRT models and for positively- and negatively-worded items.

(IRT) Item Response Theory

Poster 76: INVESTIGATING PRACTICALITY FOR BUILDING AN ITEM BANK USING CALR METHOD
Haruhiko Mitsunaga, Nagoya University

It is essential that proper equating method, such as Mean-Mean method (Marco,1977) and Stocking-Lord method (Stocking & Lord, 1983), is selected to obtain more accurate estimates when building an item bank based on Item Response Theory (IRT). These methods which estimate linear transformation coefficients are used to equate a pair of different item parameter sets. However, these method require multiple estimation procedure if more than two sets of item parameter are to be equated. Arai & Mayekawa (2011) proposed the CALR equating method which can estimate linear transformation coefficients for multiple item parameter sets. This method would be useful when a reference population is defined, and a field test is administered to establish the scale for the reference population prior to the examination in which the examinee shall be scored. CALR method enables us to equate the item parameter from the field test to other parameter sets from multiple examinations, but the accuracy of this method is yet to be extensively probed for its practicality. In this paper, the robustness of CALR method was examined through comparison with concurrent calibration methods and other pairwise equating methods. Simulation illustrated that the accuracy of CALR method was higher than the

pairwise equating method, and almost the same as concurrent calibration. CALR method may be useful when multiple test scales are equated by using a single estimation procedure.

(IRT) Item Response Theory

Poster 77: DIFFERENTIAL ITEM RESPONSE MODEL FOR THE EFFECT OF A CONTINUOUS PERSON COVARIATE

Saemi Park, The Ohio State University; Paul De Boeck, The Ohio State University

We propose a new model, differential item response model, for differential item measurement where a differential effect of a person covariate is postulated for the items of the entire of test and affects item easiness in a differential way. For the easiness of items we use a continuous person covariate that offers several advantages we introduce. We have applied DIRM to a vocabulary knowledge test using a reading comprehension score a covariate. Reading comprehension allows one to related sentences and draw connections between words within a sentence. Perhaps the effect of reading comprehension is not the same for all words and it would result in differential item measurement. Since reading comprehension ability varies over time, it can function as a time-relevant variable. The approach allows us to model growth of item easiness (item growth curve) as a function of reading comprehension although it is not a longitudinal data. The violation of measurement invariance is in general a threat to validity and comparability of measurement scores. However we believe studying differential measurement is informative of the processes underlying the acquisition of word knowledge. We found a significant and slightly quadratic differential effect of reading ability on vocabulary items. The item slopes can be partially explained through the polysemy and frequency of the words. It turned out that the strong readers perform better and more so the more polysemous and frequent the words are. The quadratic trend implies that the curves become steeper the better the reading comprehension of the students is.

(IRT) Item Response Theory

Poster 78: BETA FACTOR MODEL FOR BOUNDED CONTINUOUS RESPONSES

Javier Revuelta, Autonoma University of Madrid

This work introduces a beta factor model for continuous observed variables bounded between 0 and 1. In psychological and educational measurement these data typically consists of ratings or proportions. The traditional linear factor model is not entirely satisfactory for bounded continuous data because it can make predictions out of range and has no flexibility to capture skewness. In recent years, numerous applications of the beta distribution have been developed in the context of regression models and unidimensional item response modeling. The beta distribution is defined in the range (0, 1), allows for asymmetry present in bounded variables and assumes different distributions depending on the value of its parameters. A multiple factor model and its corresponding estimation algorithms are presented in this poster. The beta factor model is estimated using a marginal maximum likelihood/EM algorithm implemented using several integration methods: static Gauss-Hermite quadrature, adaptive GH, Monte Carlo EM and Metropolis-Hastings Ronbbins-Monro. A real data analysis is included in which an exploratory beta factor analysis applies to a scale of conservatism. The task of the examinees consists of rating twelve statements using a bounded continuous scale. Models with different number of factors were compared using likelihood-ratio chi square, AIC and BIC. The conclusion is that three dimensions are needed to explain these data, and these factors are interpreted as different aspect of conservatism.

(IRT) Item Response Theory

Poster 79: ESTIMATING CROSS-COUNTRY DIFFERENCES IN ENVIRONMENTAL ATTITUDE

Jan Urban, Charles University

Background: Cross-cultural studies of attitude are typically based on classical test theory, use verbal evaluative statements to estimate attitude levels, and rarely establish scalar measurement invariance before comparing attitude levels across countries (e.g., Franzen and Meyer, 2010; Hunter, 2004). By ignoring item difficulties of attitudinal responses and their known variation across countries (e.g., Kaiser

and Biel, 2000), such studies derive measures which are not sufficiently sensitive to extreme levels of attitude and, moreover, cannot be compared across countries (e.g., Urban, 2016). Method We use a multilevel latent regression Rasch model to estimate environmental attitude model based on 20 self-reports of environmental behavior and evaluative statements in 28 European countries (total N = 27,998). The model is estimated in the Bayesian framework in RStan. We make three alternative assumptions regarding the cross-country linking: (M1), we select items which are known to have highest cross-cultural invariance and use them for linking; (M2) we assume the the linking items vary in their difficulty across the countries around average cross-country difficulty; (M3), we assume that not only the linking items but also other items vary around their cross-country everage difficulties. Results All three models have high predictive validity. However, Bayesian fit analysis shows M2 (partial linking) fits best. This model is theoretically consistent with known cross-country variations in difficulty of environmental behaviors but it also show limits of cross-country comparison of attitude levels because variation in item difficulties confounds country-specific estimates of attitude levels.

(IRT) Item Response Theory

Poster 80: IRT ANALYSIS OF PISA 2015 ENVIRONMENTAL AWARENESS

Shuang Wang, Beijing Normal University

The improvement of environmental awareness can greatly foster civil behaviors of guaranteeing ecological environment. Using Item Response Theory and General Partial Credit Model to analyze the psychometric properties environmental awareness questionnaire with Chinese 15-year-old students from PISA 2015.Primary Component analysis(PCA) and General Partial Credit Model are adopted. The results show that environmental awareness measures one latent trait so the unidimensionality assumption is achieved. All items have moderate to high discriminations, the corresponding difficulty parameters range from -3 to 3. All items provide acceptable information.

(IRT) Item Response Theory

Poster 81: TRUSTING THURSTONE: IDEAL-POINT VERSUS DOMINANCE MODELS FOR TRUST IN SCIENCE

Samuel Wilgus, North Carolina State University; Justin Travis, North Carolina State University

The majority of attitude and belief measures are developed using the procedures of Rensis Likert and evaluated against psychometric criteria assuming a dominance response process. Relatively few contemporary measures are developed from Thurstone's (1928;1929) scaling procedures, which are congruent with an ideal-point response process. While responses to measures of many constructs (e.g., cognitive ability) are best evaluated using a dominance model, there are legitimate concerns that responses to measures of particular constructs (e.g., personality or attitudes) are more congruent with an ideal-point process. This study compares two sets of response process models, dominance and ideal-point, applied to a trust in science measure developed from Thurstone's methodology. A total of four approaches are employed using a 2 (dominance vs. ideal-point) x 2 (observed vs. model-based) design to examine psychometric properties of the trust in science scale, as well as estimating validity coefficients with political beliefs, education level, and beliefs about scientific conclusions in a convenience sample of 401 adults. Results suggest that both the ideal-point and two-parameter IRT models fit equally well in terms of overall model-data fit despite the data showing no unfolding across levels of theta, demonstrating flexibility of the ideal-point IRT model for capturing non-ideal-point response patterns. We conclude that if scientists have theoretical reasons to believe that a measure evokes an ideal-point response process, then using ideal-point IRT techniques may be more robust if a dominance response actually exists than vice-versa (e.g., using two-parameter IRT for ideal-point response process).

(IRT) Item Response Theory

Poster 82: EMPIRICAL TRYOUT OF A NEW STATISTIC FOR DETECTING TEMPORALLY INCONSISTENT RESPONDERS

Matthew Kerry, Swiss Federal Institute of Technology (ETH-Zurich)

Statistical screening of self-report data is typically conducted to support quality of analyzed responses. A new index (Dptc) was recently developed to identify temporal outliers in repeated-measure designs. Dptc's central premise is substitution of 'centered values' in the original Mahalanobis-D formula with 'individual-difference scores' between two assessment occasions. The Dptc's introduction on simulation data has limited empirical research uptake. The proposed poster reports on the Dptc's performance ¬across three empirical samples. Sample1 used a pre-post administration (n=620) of Future Time Perspective questionnaire in older adults (age>40-years). Hypothesis(1) supported Dptc's positive association with chronological age (age-related inconsistency) and hypothesis(2) (reverse score-related inconsistency) supported Dptc's first-empirical application. Sample2 examined Dptc for team-level analyses (n=24) with team-efficacy reports. Hypothesis(3) (outcome efficacy-related inconsistency) and hypothesis(4) (agreement-related process efficacy inconsistency) further supported Dptc's viability for use in mixed-levels research. Sample3 used a randomized-control trial of subjective-life expectancy (n=102) to examine Dptc's sensitivity to content responsitivity and convergence with classical screening tools (response time and instructed item response). Hypothesis(5) (treatment-related responsitivity) and hypothesis(6) (methods congruence) supported Dptc's flexible application to both observational and experimental repeated-measures designs. As a multivariate distance indicator based on raw response patterns over time (within-individual), the Dptc could become a powerful tool for strengthening the quality of within-unit datasets or repeated-measures designs. As with the original Mahalanobis-D, D2ptc is asymptotically distributed as a chi-square statistic, permitting statistical tests of significance with degrees of freedom equal to the number of test items.

(LDA) Longitudinal Data Analysis

Poster 83: DYNAESTI: DYNAMIC ABILITY ESTIMATION

Ajay Tripathi, Stanford University; Benjamin Domingue, Stanford University

Item response theory (IRT) models are widely used in educational measurement. Traditional usage has focused on estimation of a latent trait that remains static through the collection of item responses (e.g., they are accumulated at one sitting). However,this may not be the case when item responses are collected over years of education, or even over a single course lasting a few months. Over such timeframes, ability may be dynamic, complicating analyses in traditional IRT. Dynamic models for estimation of ability in longitudinal contexts have been proposed, but one weakness has been their typical reliance on parametrization. Given that we do not as of yet understand the dynamics of learning, such parametrizations seem premature. In this paper, we propose DynAEsti, an augmentation of the traditional IRT Expectation Maximization algorithm, for uses in cases where ability is highly dynamic. DynAEsti uses non-parametric techniques to capture ability curves. The performance of DynAEsti is evaluated and stress-tested through simulated examples of time series dichotomous response data wherein ability changes at high amplitude and frequency. In these tests, we achieved over 90% recovery for both the Item Response Functions (IRFs) and student ability curves. We compare recovery in the dynamic case to a scenario in which items are well-behaved (e.g., they follow the Rasch model) and abilities are static. As recovery is only marginally improved in this static case, we argue that recovery in the dynamic case is nearly optimal.

(LDA) Longitudinal Data Analysis

Poster 84: IRT VERTICAL SCALES WITH MIXED FORMAT TESTS

Yu-Lim Kang, Yonsei University; Guemin Lee, Yonsei University; Jung-A Han, Yonsei University & Korea Institute for Curriculum and Evaluation

Vertical scaling is a series of procedures that convert test scores from different grades into a common scale. The vertical scale developed from vertical scaling provides information on student growth (Kolen & Brennan, 2014). The factors that may affect the characteristics of vertical scales need to be carefully investigated (Tong & Kolen, 2010). In most large scale assessment programs, mixed format tests are

widely used due to the advantages covering a broad range of contents and measuring higher level thinking processes. However, relatively few studies were conducted to examine IRT vertical scaling with mixed format tests (Kirkpatrick, 2005; Meng, 2007; Moore, 2015). In mixed format tests, multidimensionality might be present due to different item format effects. When applying unidimensional vertical scaling method to the mixed format tests, the characteristics of vertical scales need to be explored. This study was designed to investigate the growth pattern and scale variability of vertical scales with mixed format tests. Simulation techniques were implemented to achieve research objectives with three factors. Simulation study is conducted under degree of multidimensionality (correlation between MC and CR traits), configuration of common items (MC only, CR only, and MC+CR), and difference of ability distribution among grade levels.

(MDS) Multidimensional Scaling

Poster 85: AN INVESTIGATION ON THE EFFECTS OF INTEGRATED TEST DIMENSIONS ON EQUATING
Feifei Li, Educational Testing Service

Tests that integrate the skill sets or content areas are more often used for academic achievement assessments, for example, science tests that are designed under the guidance of Next Generation Science Standards (NGSS) to measure students' integrated science proficiency. Given the limited amount of testing time, students cannot be tested on all items, but are assigned with item blocks that measure the general proficiency on all content areas and also item blocks that focus on certain single areas. Concurrent calibration will be run so that the item parameter estimates are placed on common scales. As such, it is possible that the weight on each dimension varies for students that take different combination of item blocks. Another consideration is that the tests have been specifically designed to measure content areas related to course curriculum. However, students may take the tests at different points in their coursework. In that case, there is a need to investigate multidimensional item response theory (IRT) calibration and equating for the afore-mentioned tests and using the population as described above, in particular, when the a test combination with higher weight in an area is administered to students who are weak in that area. The current study is intended to examine four factors through simulation: (1) the correlations between dimensions; (2) ability distribution in the equating sample; (3) random assignment; (4) sample size. This study will provide a better understanding of the possible effects of the multidimensional integrated tests with block design on calibration and equating.

(MDS) Multidimensional Scaling

Poster 86: SECOND-ORDER MIRT MODEL FOR EQUATING SUBSET SCORES
Youkyoung Oh, Yonsei University; Guemin Lee, Yonsei University

Many large-scale testing programs report subtest scores as well as a total test score to provide diagnostic information in different content areas (Haberman&Sinharay, 2010). A total test often intended to measure a single general proficiency, it is very likely that different subtests measure somewhat different, yet highly correlated, proficiencies (Tate, 2010). However, in unidimensional item response theory (UIRT) framework, correlation information between subtest is typically ignored (de la Torre & Song, 2009; Rigimen, Jeon, von Davier, Rabe-Hesketh, 2014). In this context, Yao (2010) and de la Torre & Song (2009) proposed second-order multidimensional item response theory (MIRT) model that the correlations between subtest-specific factors are modeled through a second-order factor.

(MDS) Multidimensional Scaling

Poster 87: IMPUTATION OF MULTILEVEL MISSING ITEM RESPONSE DATA WITH FCS
Holmes Finch, Ball State University; Julianne Edwards, Azusa Pacific University

Frequently, educational/psychological measurement data are collected in a multilevel framework, such as students nested within schools. During such data collection, it is common for data to be missing. To address missing data, a number of imputation techniques have been proposed in the context of Item Response Theory (IRT). Though a number of approaches are effective for single level data, the literature

would suggest that they may not be appropriate for multilevel data (Schafer, 2001), or may only work in specific situations (Enders, Mistler, & Keller, 2016). Recently, a fully conditional specification (FCS) method for missing value imputation with multilevel data was developed (Enders et al., 2017). This multilevel FCS approach has been demonstrated to be effective for two-level regression models, yielding data that contained low parameter estimation bias for a random slopes multilevel regression model. Multilevel FCS imputation has not been investigated in the context of categorical variables, such as item response data. Thus, the purpose of this simulation study is to extend earlier work by applying multilevel FCS to missing dichotomous item responses in a multilevel IRT framework. The manipulated conditions in this study include number of clusters, sample size per cluster, number of items, proportion of missing data, and intraclass correlation value. Methods for dealing with missing data include multilevel FCS, a naïve FCS approach assuming single level data, and listwise deletion. Outcomes of interest include item parameter estimation bias, coverage rates, and root-mean square error, and bias, coverage rates, and root-mean square error for variance components.

(MIS) Missing Data

Poster 88: THE IMPACTS AND TREATMENTS OF MISSING DATA IN CDM

Hueying Tzou, National University of Tainan; Ya-Huei Yang, National University of Tainan

The low birthrate has affected Taiwan education continuously. To solve the problem caused by low birthrate, the Ministry of Education in Taiwan reduced the numbers of students in each class to provide elaborated instruction. Also, the remedial teaching programs were promoted for low achievement students to shrink the learning gap and ensure their basic competency. This made cognitive diagnostic models (CDMs) more attractive in Taiwan education since CDMs could provide refined information on students' strength and weakness. However, while in the application of CDMs, the problem of missing data is unavoidable. Unfortunately, the impact of missing data in CDMs is still in an undeveloped state. Therefore, the purposes of this study are: (i) to explore the impact of different missing mechanisms (missing completely at random, missing at random, missing not at random) on the item parameters and the correct classification rates for examinees; (ii) to explore the impact of different missing data treatments (listwise deletion, zero imputation, multiple imputation) on the item parameters and the correct classification rates for examinees. To be close to the real condition, the empirical Q-matrix and the item parameters from the fraction subtraction data of Tatsuoka (1984) are used for generating the simulation data sets. Except for the simulation studies, the study will also compare the commons and the differences between the simulated and empirical data.

(MIS) Missing Data

Poster 89: MEASURES OF MODEL FIT FOR LONGITUDINAL MULTILEVEL MODELS

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

Linear models are commonly evaluated on the basis of measures of fit (such as R-squared), and it is generally desirable that such measures possess certain properties such as: boundedness, monotonicity, linear invariance, and intuitive interpretability (e.g., Kvalseth, 1985). For multilevel models, finding such measures has proven to be a challenge, and in this simulation study we evaluate the utility of several proposed measures of fit for longitudinal multilevel models. The measures evaluated include those proposed by Jaeger, Edwards, Das, and Sen (2017); McFadden (1974); Nakagawa and Schielzeth (2013); Raudenbush and Bryk (2002); and Snijders and Bosker (2012). The simulation study involves models of outcomes measured on either 4 or 8 occasions for samples of either 30 or 200 individuals; thus measurement occasions (level-1 units) are considered to be nested within individuals (level-2 units). Additional factors varied in the simulations include different levels of model complexity (i.e., number of predictors at level-1 and level-2), size of the predictor coefficients, predictor collinearity levels, and the (mis)specification of random components. The results are used to evaluate the measures of fit in terms of (1) estimation (by comparing sample and pseudo-population results); and (2) adherence to the properties of boundedness, monotonicity, and intuitive interpretability (by comparing results across models of different complexities). The results from this study will be used to inform and provide recommendations to researchers who wish to report a measure of fit for longitudinal multilevel models.

(MLM) Multilevel/Hierarchical/Mixed Models

Poster 90: SCHOOL SOCIOECONOMIC STATUS MODERATE STUDENTS' MATH SELF-EFFICACY ON MATH ACHIEVEMENT

Ruiyan Gao, Beijing Normal University; Xiaojian Sun, Beijing Normal University; Yang Tao, Beijing Normal University

The study aims to investigate how does school socioeconomic status (school SES) moderate students' math self-efficacy on their mathematics performance. By adopting two-level models to analyze Shanghai, China (N=5177) and the U.S. (N=4978) data from PISA 2012, the results show that although both school SES and individual math self-efficacy could positively predict students' mathematics achievement, there is a heterogeneous pattern for these countries. For Shanghai, China, school SES negatively moderated students' math self-efficacy on mathematics achievement, which means math self-efficacy could explain more variation in low SES schools than in high SES schools for students' math achievement. However, for the U.S., school SES positively moderated students' math self-efficacy on mathematics achievement, which means math self-efficacy could explain more variation in high SES schools than in low SES schools for students' math achievement. The findings suggest that there may have a mathematics self-efficacy compensation for low SES schools in Shanghai, China, and there may have a mathematics self-efficacy enhancement for high SES schools in the U.S. The cultural differences between two nations could be used to explain this phenomenon, and the different school environment and school regulation between western and eastern were also considered.

(MLM) Multilevel/Hierarchical/Mixed Models

Poster 91: THE EFFECT OF INCIDENTAL SECOND-LEVEL DEPENDENCE IN MULTILEVEL MODELS

Weimeng Wang, University of Maryland, College Park; Manqian Liao, University of Maryland, College Park; Laura Stapleton, University of Maryland, College Park

Unmodeled between-level dependence can affect parameter estimates and inference in MLMs. Between-level dependence can be quite common in educational and psychological research especially when the level of clustering is not part of the sampling design. In many national educational data collection programs from the National Center for Education Statistics, schools are often used as the first-stage selection process and students are selected from schools, but the classroom level is not part of the design. However, given a limited number of classrooms within a school, it is highly likely some of these randomly selected students share the same classroom context. The incidental dependence of students within the same classrooms violates the assumption of independence of residuals at each level and could potentially influence the standard errors of the estimates. In the current study, we investigate the influence of incidental level-2 dependence on the parameter estimates including the fixed effects and the random variance components, and their corresponding standard error estimates. This study conducts a simulation study to examine the performance of a 2-level model and a 3-level model of modeling the incidental second level dependence. A motivating educational example using data sets of Early Childhood Longitudinal Study-Kindergarten of 1998-99 (Tourangeau et al., 2009) also shows how different ways of modeling second level dependence can affect the parameters.

(MLM) Multilevel/Hierarchical/Mixed Models

Poster 92: A FRAMEWORK OF EQUATING IN COGNITIVE DIAGNOSTIC MEASUREMENT

Liping Yang, Beijing Normal University

Compared with traditional psychometrics scoring on a continuum, cognitive diagnostic measurement (CDM) provides decisions on multiple characteristics of an individual. However, many of the modeling advances have not been fully realized in practice, one of the reasons is the lack of method to compare CDM scores obtained from multiple test forms. Test equating refers to measuring the same psychological traits of multiple test form of test score or item parameters conversion, corresponding with indicators comparable process. The current equivalent study focused on the framework of the classical measurement theory (CTT) and the item response theory (IRT), there is little research on the equivalent of CDM. Furthermore, the common item (or person) equating design has obvious limitations in practice, the methodologic innovation of equating need to be developed. This research aims to (1)

propose an equating framework of CDM that builds upon the general equating practice as discussed in Kolen and Brennan (2014), (2) use anchor attribute non-equivalent group design, and develop attribute characteristic curve equating method, (3) build a set of link and transformation formula among equivalent coefficient under models of CTT, IRT and CDM in separate calibration, (4) Four types of scores for individual will be equated in this study, including attribute profile (the mastery/non-mastery decision on each attribute), a probability of mastery on each attribute, ability parameter in IRT and the raw score of test, (5) discrimination parameters and difficulty parameters, both on attribute level and item level will also be transformed into a same scale.

(MLM) Multilevel/Hierarchical/Mixed Models

Poster 93: LONGITUDINAL DATA ANALYSES WITH MULTIVARIATE LATENT GROWTH MODEL
Jung-A Han, Yonsei University & Korea Institute for Curriculum and Evaluation; Guemin Lee, Yonsei University; Yu-Lim Kang, Yonsei Universtiy

In the longitudinal data analysis, the latent growth model can be used to identify the relationship between the significance of the individual differences of the changes, the factors directly or indirectly affecting the individual differences, and the changes in one variable and the other. In this structure, it is assumed that the influence of variable that changes with time is different at each measurement point. However, if the time-varying variable is interested in the impact of the change process, or if the time-varying variable itself undergoes a process of systematic change or growth, this structure will not fully utilize the information it holds. Therefore, it does not lead to an accurate interpretation of the relationship between two variables. In this case, after identifying the change model of all related variables and constructing the multivariate latent growth model by combining the two models, the information of the variable can be used effectively. Therefore, in this study, we try to identify the effect of change of learning activities on the change of life satisfaction in adolescents by applying multivariate latent growth model. The analysis procedure is as follows. First, the changes of learning activities and life satisfaction were explored separately in four-year data. Second, by combining the two models it was searched for the relationship between changes in the two variables. At this time, negative parenting style, their socio-economic background, and self-esteem variables were added to the model as control variables.

(MVA) Multivariate Analysis

Poster 94: POMPOM: PERSON-ORIENTED METHOD AND PERTURBATION ON THE MODEL
Xiao Yang, Pennsylvania State University

Lifespan developmental theories view persons as dynamic systems, with self-organization among multiple levels of analysis (e.g., biological, psychological and social). Recent work has started to explore the possibility of merging multivariate time-series methods with network analysis to describe the individual differences in self-organization (e.g., emotion, physiological response, psychopathological symptoms), which is usually summarized in network metrics. We forward a new network metric – impulse response analysis metric (iRAM) – that quantifies regulatory efficiency. By perturbing the network inferred by a person-oriented method (unified Structural Equation Modeling), iRAM is computed as the duration from perturbation to equilibrium. Conceptually, iRAM is inversely associated with regulation efficiency (e.g., longer iRAM indicates lower efficiency). iRAM can also differentiate excitatory or inhibitory feedback loops, which are highly relevant for efficient regulation. An accompanying R package pompom (https://cran.r-project.org/web/packages/pompom/index.html) has been developed to enhance the accessibility of this hybrid method. Using a hypothetical multivariate example, we demonstrate the steps of the hybrid method and associated R functions, including model specification of uSEM, extraction of model fit statistics, computation of perturbation, bootstrapped calculation of iRAM, and visualization of network and dynamic influence. Findings highlight the utility of iRAM to quantify regulatory efficiency in a multivariate system, and more broadly, the utility of using a person-oriented method and the perturbation approach to articulate and test hypotheses about individuals as complex, high-dimensional dynamic systems.

(NET) Network Analysis

Poster 95: KNOWGENE SCALE: A MIXED METHODS APPROACH FOR ASSESSING RESPONSE PROCESSES

Daniel Gundersen, Dana-Farber Cancer Institute; Jill E. Stopfer, Dana-Farber Cancer Institute; Anu Chittenden, Dana-Farber Cancer Institute; Meghan Underhill, Dana-Farber Cancer Institute

Developing knowledge assessments for multigene panel testing for cancer risk is challenging, but needed for clinical practice and research. Items must be non-threating despite technical terminology during a time of potentially high cognitive load, and meaningfully inform clinicians for targeting education for under- vs. mis-informed patients. The KnowGene instrument was developed using a mixed methods approach to address these challenges. Five national subject area experts, including researchers and clinicians, and 4 patient advocates reviewed candidate items for content validity and acceptability. Cognitive testing was used to evaluate comprehension and response process. A 2 parameter logistic-nested logit (2PL-NLM) Item Response Theory (IRT) model was used to estimate the probability of a correct response and, conditional on incorrect response, distractor category selection as a function of ability level. For locally dependent item-pairs, the item with greater discrimination was retained unless it compromised content validity or its location parameter was in a range with good coverage. The qualitative assessment determined offering a "don't know" category would make the instrument less intimidating, would allow clinicians to differentiate patients on educational need, and suggested a hierarchical evaluation of response categories. Item non-response was <1% for all items, suggesting the instrument was not intimidating. A unidimensional 2PL-NLM found 18 items with high information for lower ability levels. Greater ability level was associated with greater probability of selecting a distractor category over "don't know." Qualitative approaches informed strategies of making items less intimidating, permit clinicians to target education, and selection of appropriate IRT approach.

(PRO) Patient-Reported Outcomes

Poster 96: CREATING MISFIT FOR MOMENT STRUCTURES GIVEN Θ AND F_ML VALUES

Keke Lai, University of California, Merced

To understand how SEM methods perform in practice where models always have misfit, simulation studies often involve incorrect models. To create a wrong model, traditionally one specifies a perfect model first and then removes some paths. This approach becomes difficult or even impossible to implement in moment structure analysis, and fails to control the amounts of misfit separately and precisely for the mean and covariance parts. Most importantly, this approach assumes a perfect model exists and wrong models can eventually be made perfect, whereas in practice models are all implausible if taken literally and at best provide approximations of the real world. To improve the traditional approach, we propose a more realistic and flexible way to create model misfit for multiple group moment structure analysis. Given (a) the model $\mu(\cdot)$ and $\Sigma(\cdot)$, (b) population model parameters θ_0, and (c) F_1 and F_2 specified by the researcher, our method creates μ^* and Σ^* to simultaneously satisfy (a) $\theta_0 = \text{argmin } F[\mu^*, \Sigma^*; \mu(\cdot), \Sigma(\cdot)]$, (b) the mean structure's misfit equals F_1, and (c) the covariance structure's misfit equals F_2.

(RES) Resampling and Simulation Techniques

Poster 97: SAMPLE SIZE AND STATISTICAL POWER FOR SEM: A SIMULATION STUDY

Ning Jiang, University of South Carolina; Jin Liu, University of South Carolina; Dexin Shi, University of South Carolina; Christine DiStefano, University of South Carolina

When researchers start their studies using structural equation modeling (SEM), one of critical steps is to determine a minimum sample size to reach an adequate statistical power (e.g., 0.8). However, rules of thumb are not always accurate, and determining sample size and power challenges most researchers. Further, many studies investigating power in the SEM field have considered continuous data. As ordinal data are often analyzed, further research involving discrete data are needed. This study aims to use Monte Carlo method to examine minimum sample size to reach adequate statistical power (0.8) when ordered data are analyzed. Conditions such as the number of categories, magnitude of factor loadings, misspecification levels, and item distribution are varied. Mplus (v. 8) is used with the robust estimator of WLSMV (Mean- and Variance-adjusted Weighted Least Square), and one thousand replications are run

for each design cell. Preliminary results are consistent with the previous studies which indicate that sample size and statistical power are impacted by magnitude of factor loadings, number of data categories, and data normality. The findings are useful for researchers and practitioners to use Monte Carlo method to determine sample size and statistical power when they choose their own SEM models.

(SEM) Structural Equation Modeling

Poster 98: BAYESIAN APPROACHES TO DETECT INTERPRETATIONAL CONFOUNDING IN FORMATIVE MEASUREMENT MODELS

Houston F. Lester, Center for Innovations in Quality, Effectiveness, and Safety, Michael E. DeBakey VA Medical Center, Houston; James A. Bovaird, University of Nebraska-Lincoln; Nebraska Academy for Methodology, Analytics, & Psychometrics

This study discusses a novel approach for assessing the validity of measurement models containing both causal-formative and effect indicators. Formative measurement models have been criticized because some researchers believe these models are inherently subject to interpretational confounding (i.e., a mismatch between the theoretical and empirical functioning of a construct; Burt, 1976). Prior research has demonstrated that formative measurement models are not inherently subject to interpretational confounding (Bainter & Bollen, 2014); however, closer inspection of that research reveals how easily interpretational confounding can occur if researchers adhere to common SEM practices (i.e., determining that a model is adequate if the model fits with adequate effect sizes). To address this limitation, we employ Bayesian informative hypothesis testing (Van de Schoot, Hoijtink, Hallquist, & Boelen, 2012) to test a priori hypotheses regarding the rank-ordering of the causal-formative indicator coefficient magnitudes. Support for this rank ordering provides evidence that the empirical and theoretical behavior of the construct match. We conducted a simulation study to assess how well this method can detect the presence/absence of interpretational confounding and provide an empirical example. We manipulated factors that determine the presence of interpretational confounding as well as factors that may affect detection in practice (i.e., multicollinearity, number of effect indicators, and sample size). The results revealed that the Bayesian informed hypotheses could detect interpretational confounding as well as support the correct hypothesis. Thus, Bayesian informed hypothesis testing is a tool to be utilized in situations where conventional SEM practice does not detect interpretational confounding.

(SEM) Structural Equation Modeling

Poster 99: STATISTICAL ESTIMATION OF SEM MODELS WITH MIXED-SCALE OBSERVED VARIABLES

Cheng-Hsien Li, National Sun Yat-sen University; Sen-Kai Yang, National Sun Yat-sen University

In the educational, social, and behavioral sciences, observed variables of mixed scale types (i.e. both continuous and ordinal observed variables) have been applied in the structural equation modeling framework. However, there is limited knowledge regarding the impact of mixed-scale observed variables on the performance of existing estimation methods. A Monte Carlo simulation study was carried out to examine the performance of the two estimation methods with robust corrections, maximum likelihood (ML) and diagonally weighted least squares (DWLS), on parameter estimates, standard errors, and chi-square statistics in structural equation models. Conditions varying the number of ordinal observed variables' categories (5, 6, and 7), ordinal observed distributions (symmetry and slight asymmetry), population model specifications (measurement and structural equation models), and sample size (n = 200, 500, and 1,000) were examined. Data generation and analysis were performed with Mplus 8. Results reveal that (1) DWLS yields more accurate factor loading estimates of ordinal observed variables than ML whereas DWLS and ML produce comparable factor loading estimates of continuous observed variables; (2) inter-factor correlation and structural coefficient estimates under DWLS and ML outperform equally well in nearly all conditions; (3) robust standard errors of parameter estimates obtained with ML are slightly more accurate than those produced by DWLS in almost every condition, unless a larger sample is used (i.e., n = 1,000); and (4) DWLS is relatively superior to ML in controlling for Type I error rates; however, the superiority appears attenuated with increasing sample sizes.

(SEM) Structural Equation Modeling

Poster 100: SAMPLE SIZE REQUIREMENTS FOR STRUCTURAL EQUATION MODEL SELECTION
Hao Luo, The University of Hong Kong; Björn Andersson, Centre for Educational Measurement, University of Oslo

A typical application of structural equation modeling in social and behavioral studies is to evaluate a theoretical hypothesis by selecting the best-fit model among several candidate models using a selection criterion. Determining the sample size required for identifying the true model from alternative ones is challenging since sample size requirements change as a function of variable type, model properties, and choice of estimation method. Although several rules-of-thumb exist for advising applied researchers, they are not model-specific and may lead to incorrect model selection. This study uses Monte Carlo simulation to estimate the sample size requirement for selecting the true model from alternative models with different degrees of misspecification. The effect of the number of latent and observed variables, the size of factor loadings and path coefficients, and the pattern of missing values is investigated systematically. We will also examine the effect of the estimation method used and the types of variables in the model. The empirical relationships between sample size requirements and a range of choices for parameter values will be explored to provide practitioners more specific guidance about what sample size is required for a specific model.

(SEM) Structural Equation Modeling

Poster 101: DEVIATIONS FROM NORMALITY: EFFECTS ON GROWTH CURVE MODELS
Catarina Marques, Instituto Universitário de Lisboa (ISCTE-IUL); Maria de Fátima Salgueiro, Instituto Universitário de Lisboa (ISCTE-IUL); Paula C.R. Vicente, Business Research Unit (BRU-IUL), ISCTE-IUL, Lisboa

Latent growth curve models (LGCM) became in recent years a very popular technique for longitudinal data analysis: they allow individuals to have distinct growth trajectories over time. These patterns of change are summarized in relatively few parameters: the means and variances of the random effects (random intercept and random slope), as well as the covariance between intercept and slope (Bollen & Curran, 2006). Although the specified model structure imposes normality assumptions, the data analyst often faces data deviations from normality, implying mild, moderate or even severe values for skewness and or kurtosis. A traditional approach for generating data that deviates from the normal distribution was proposed by Vale-Maurelli (1983). Recently, Foldnes and Olsson (2016) proposed the independent generator transform approach to generate multivariate non-normal distributed data. Following this new approach, in the current paper a Monte Carlo simulation study was conducted in R, using lavaan, in order to investigate the effect of observed data deviations from normality on the standard errors and goodness of fit indices. LGCMs with unconditional linear growth are considered. Three and four time points, and sample sizes ranging from 50 to 500 observations are used. The impacts of such deviations on parameter estimates, standard error and fit measures are discussed.

(SEM) Structural Equation Modeling

Poster 102: SHAPING STATE EDUCATION POLICY: ROLL-CALL VOTING ANALYSIS ON PROPOSITION 58 IN CALIFORNIA
Jisung Yoo, University of Georgia

Despite the considerable literature on policymaking in the U.S. Congress, little research has examined educational policymaking process at the state and local district levels. Attempting to help fill this gap, this study examines the specific case of California State Assembly and Senate members' voting decisions on Proposition 58, a policy allowing schools to utilize multiple language programs, such as bilingual education, thereby replacing the previous requirement of English-only education for English learners. Structural equation modeling was employed to examine the influence of various factors on the voting decisions of California State Assembly members and Senators. The analysis revealed statistically significant directional relationships among such factors as characteristics of Assembly members including education level, race, ideology, and political party affiliation, and characteristics of district constituents. The findings reveal how educational policy is shaped in the state legislature and guide policymakers in developing educational policy to achieve equitable educational opportunities for all

students including English learners. Keywords: policymaking, educational policy, English language learners, California Assembly voting decision

(SEM) Structural Equation Modeling

Poster 103: PREDICTING TURNOVER INTENTIONS USING PSYCHOMETRICS AND MACHINE LEARNING
Igor Menezes, University of Lincoln; Ana Cristina Menezes, Federal University of Bahia; Kai Ruggeri, Columbia University; Elton Moraes, Korn Ferry Kay Group

Turnover intentions can be thought of as the conscious will to leave a current organization, which is deemed amongst the most significant challenges companies face. Although a few independent studies have been conducted in order to predict turnover intentions, no research has yet been carried out in order to integrate different constructs into a single instrument. Hence, this study aimed to develop a machine learning based application to identify which variables play a significant role in the prediction of turnover intentions, including experience with current organization (role clarity, job satisfaction, job credit, role conflict, value match), decision making style (resilience, emotion and cognition), organisational openness to experience (personal experience and resistance to change) and personality. Data were collected from 408 Twitter users, who answered 264 items. Graded Response Model (GRM) was used for preprocessing the dependent variable and calculate the EAP factor scores, which were then categorised into 'high intention' and 'low intention', with a threshold at 80th percentile. GRM was once again deployed, but this time as a procedure for feature selection. This reduced down the total number of items to 60. Finally, Extreme Gradient Boosting (XGBoost) was used as a classification model for the prediction of turnover intentions. The accuracy of the final model was 87.7% and the features were listed by importance according to three parameters: gain, cover and frequency. The findings of this research may provide insights into the prediction of turnover intentions and further contribute to the development of this topic and its measurement.

(SML) Statistical and Machine Learning

Poster 104: USING NATURAL LANGUAGE PROCESSING IN AUTOMATIC ITEM GENERATION
Tingting Sun, Beijing Normal University, Collaborative Innovation Center of Assessment toward Basic Education Quality

Test item plays a fundamental role in educational and psychological assessment. Research on Automatic Item Generation (AIG) has been evolving over the years, which focuses on investigating how to design a item model as radicals to generate similar items and create generating program or software as item generators (Irvine et al.,2002; Bejar et al.,2003;Embretson et al.,2007; Gierl et al.,2009,2012,2013,2015;Lathrop et al.,2017). Based on this, the objective of this study is to use the Natural Language Processing (NLP) technology to generate items automatically, taking the content of “Numeric and Algebra” in grade 4 and 8 as an Example. Data from a national assessment program in China were used in this study, and take Mathematics textbooks and as a corpus. The Practical Implications of this study is to apply AIG in large-scale achievement assessment and validate the use of NLP methods, which can grately decrease the cost and time of item writing.

(SML) Statistical and Machine Learning

Poster 105: EXPLORATORY ANALYSES OF DATA FROM A COLLABORATIVE PROBLEM SOLVING ASSESSMENT
Junyan Yao, New York University; Kaushik Mohan, New York University; Yoav Bergner, New York University; Peter Halpin, New York University

Technology-based assessments that involve collaboration among students offer many sources of process data, although it remains unclear which aspects of these data are most meaningful for making inferences about students’ collaborative skills. Identification of informative features is a crucial step in helping students improve their collaborative skills. Recent research has focused largely on theory-based rubrics for coding of process data (e.g., text from chat dialogues, click-stream data), but many reliability and validity issues arise in the application of such rubrics. In this research, we take a more data-driven

approach to the problem. Drawing on real data in which dyads collaborate via online chat to answer twelfth-grade mathematics items, we focus on features of chat and click-stream that can be extracted automatically. Examples include the extent to which students' dialogue includes content from assessment materials or from other established lexicons; the sequential dependence between student chat and student response behavior; the extent of temporal synchronization in students' response times and viewing times. The main component of this research is the exploration and interpretation of such features, resulting in a rich, multidimensional description of collaborative process data. We also consider team composition factors and group outcome data on the math assessment. The second main component of the research is to apply unsupervised machine learning techniques to infer whether different classes of dyads and individuals can be identified using features from processes, team composition, and outcomes.

(SML) Statistical and Machine Learning

Poster 106: INVESTIGATING VALIDITY AND PRECISION WHEN SHORTENING A SPEEDED TEST

Daniel Adams, University of Wisconsin, Madison; Rich Feinberg, National Board of Medical Examiners; Peter Baldwin, National Board of Medical Examiners

Speededness occurs when examinees do not have sufficient time to fully consider all items. One way to reduce speededness is to increase testing time; however, when this approach is infeasible due to cost increases or other reasons, another option is to reduce the number of items while keeping the testing time unchanged. Removing items increases time per item but decreases precision. This study proposes a method for investigating changes in score validity and precision under these conditions. Two years of item response data were obtained from a high-stakes examination. In both years the test comprised seven one-hour blocks; however, in year two, the number of items per block was reduced from 44 to 40. In addition to total test score, each examinee was given two additional scores: a speeded score based on the final 5 items within each block and a random score also based on 5 items from each block but selected at random. Using total score as the criterion, the expected deviation and root-mean-square deviation then were calculated for each of these two scores for each year. (In the case of the random scores, expected deviations and root-mean-square deviations were additionally averaged across 25 replications.) It was found that differences between speeded and random expected deviations and speeded and random root-mean-square deviations declined across years. These finding suggests that the reduction in the number of test questions per block may have had the intended effect of reducing systematic and random error due to speededness.

(VAL) Validity and Reliability

Poster 107: APPLICATION OF A METHODOLOGY FOR TEST ADAPTATION IN EDUCATIONAL CONTEXTS

Joaquin Caso Niebla, Universidad Autonoma De Baja California; Carlos David Diaz Lopez, Universidad Complutense De Madrid; Coral Gonzalez Barbera, Universidad Complutense de Madrid

The aim of this study was to develop and apply a methodology for tests adaptation in educational contexts attending to the guidelines, standards, and recommendations from different organizations (AERA, APA & NCME, 2014; ITC, 2010) and authors (Hambleton, 2005; Muñiz & Hambleton, 2013; Van de Vijver & Tanzer, 2004; Zumbo, 1999). We approach the test adaptation in two phases and six stages: (a) the rational-empirical phase comprises test translation, linguistic validation, content validation and pilot testing, and (b) the psychometrics phase considers test application and test for bias and equivalence. This methodology considers psychological, psychometric, cultural and linguistic aspects as part of the process of test adaptation. We document all the evidence obtained throughout all these stages in the adaptation of the School Violence Scale, developed in Spain (Ortega, Del Rey & Casas), into de Mexican context. A total of 10 judges and 927 elementary and high school students participated. Both the confirmatory factor analysis and the differential item function analysis (in the test for bias and equivalence stage), confirm that the adapted version of the scale measures the same dimensions as the original version and that its items are free of bias. The collected evidence from each stage supports both the use of the test in the new contexts and the need to use this kind of approach that allows us to deal

with the cultural differences (even between countries that speak the same language) in the explanation of educational outcomes.

(VAL) Validity and Reliability

Poster 108: RUMOR VALIDITY SCALE

Joshua Chiroma Gandi, University of Jos; Wukatda Beben Wokji, Plateau AIDS Control Agency, Nigeria; Dauda Akwai Saleh, Plateau State University, Nigeria; Paul Samani Wai, Plateau State University, Nigeria

Rumor refers to unverified information being circulated among people to make sense of an unclear situation or to manage any threat or potential threat (Matsumoto, 2009, p.451). It has been observed that the more the integrity of the source, the more reliable the rumor would be. But not all that is reliable would be summarily adjudged as valid. There's need for empirical validity assessment of rumors in any case. However, despite the fact that psychological scales are indispensable for assessment and data collection, there has been a dearth of construct-relevant scale for rumor validity assessment. The present study was designed to develop a construct-relevant scale for assessing rumor validity. A cross-sectional design was adopted and a sample of 570 randomly selected respondents provided the validation data. The validity of the scale was determined based on the model emphasized by Coaley (2010, p.132) which include content-related validity, criterion-related validity, and construct-related validity. Content validity index, item-total statistics, exploratory factor analysis, confirmatory factor analysis, Pearson's correlation and multiple regressions were used. The content validity indexes of 0.92, 0.90 and 0.85 which followed item generation, pretesting and pilot-testing stages attests to the suitability of the items as 'construct-relevant'. Cronbach alpha of 0.88 corroborated by split-half parts 1 (0.96) and 2 (0.93) indicated the scale's reliability. Three extracted factors with their respective item membership of 10, 8 and 7 accounted for 92% of the total scale variance. The rumor validity scale is a 25-item scale suitable for adaptation at various settings and populations.

(VAL) Validity and Reliability

Poster 109: VALIDATION OF INTEGRATED PALLIATIVE OUTCOME SCALE IN CANCER PATIENTS

Mevhibe Hodjaoglou, Eastern Mediterranean University

Palliative care aims to improve the quality of life of patients and their families and reduce suffering from life-threatening illness. In assessing palliative care efficacy, patient reported outcomes reflecting patient perspectives of their care and needs are critical. Such a tool is absent in the Turkish speaking community in Cyprus. The aim of this study was to translate, cross-culturally adapt and validate the Integrated Palliative care Outcome Scale (IPOS) Patient one-week version to the cancer care context of the Turkish speaking community in Cyprus. Turkish version of the IPOS was produced and tested in two phases. The first phase involved verification of conceptual equivalence through literature review, professional interviews, and patient focus groups, multiple forward and blind backward translations, comparisons of the translated versions, pilot testing of the pre-final version, and POS Development Team review. In the second phase 200 patients living with cancer completed EQ-5D and IPOS at time one. A subset of 100 participants completed IPOS a second time one week later. Classical Testing Theory and Item Response Theory was used to evaluate the internal consistency reliability, stability reliability (test-retest reliability), homogeneity, construct validity (convergent and divergent), criterion-related validity, factor structure of the instrument (dimensionality) and model fit were evaluated. In this study, we translated, culturally adapted and psychometrically tested IPOS Patient one-week version among Turkish speaking Cypriots living with cancer in Cyprus. The study findings indicate that the Turkish IPOS is a valid and reliable measure of palliative care outcomes with cancer patients.

(VAL) Validity and Reliability

Poster 110: COMPARISON OF STANDARD SETTING METHODS FOR ASSESSMENTS IN DENTAL EDUCATION

Muhammad Naveed Khalid, Cardiff University; Sheila Oliver, Cardiff University

Standard setting is an essential process in defining competency in dental education. Different standard setting methods tend to lead to different cut-off points which have direct consequences for the individual students. One commonly used method is the Angoff in which a panel of experts estimate the percentage of borderline students predicted to correctly answer each question in an examination. This method is costly, time consuming and relies on the assumption that the panel can accurately define the borderline student. Recently the Cohen method has been developed and subsequently modified, to overcome these disadvantages. We also examined the relative method, we took the mean of the score distribution for the group as a reference, then picked a point below that mean as the passing mark. Our aim was to compare the standard set using our current method of Angoff to relative method, Cohen and modified Cohen methods, to inform future standard setting practices. Standard setting methods were applied to historical data for written examinations across assessments of the BDS programme. Data included cohort sizes of 25-85 students per year. The procedures yielded inconsistent results. The Angoff and relative (Mean and 1SD) procedures gave similar results; however, the Cohen and Modified Cohen gave divergent results. The Angoff procedure yielded results reasonable, defensible and reliable enough to use in decision making for a high-stakes examination. Further investigation of other procedures is needed because we have examined one year data and findings are inconclusive.

(VAL) Validity and Reliability

Poster 111: THE LATENT STATE-TRAIT-MULTIMETHOD-CFA APPROACH FOR THE EVALUATION OF CONSTRUCT VALIDITY

Tuulia Ortner, University of Salzburg; Michael Eid, Free University of Berlin; Tobias Koch, Leuphana University of Lüneburg

Early approaches for the analysis of construct validity based on MTMM-designs postulate several criteria as indicators for validity, for example, sufficiently large correlation coefficients if the same constructs are measured by different methods. The application of these criteria is difficult as there has been no statistical test whether or not the criteria are fulfilled in empirical applications. Furthermore, the approach does not account for temporal effects, and, problems may occur if the included measures differ in their reliability. We present data examining the convergent and discriminant validity of Objective Personality Tests, IATs and self-report measures of four different traits. Objective Personality Tests (OPTs) that aim to assess personality characteristics in computerized and standardized settings without relying on self-report measures have been rarely studied with reference to their construct validity. Especially broad studies investigating convergent and discriminant validity with respect to other methods of personality assessment (indirect measures, self-report measures) have been seldom conducted so far. The convergent and discriminant validity of OPTS was assessed on trait (stable) and state (momentary) level by using the multimethod latent-state-trait (MM-LST) model developed by Courvoisier et al. (2006). The MM-LST model allows separating different sources of variances: stable and momentary trait influences, stable and momentary method influences and measurement error influences. In total, this study incorporated data from 300 students assessed on three different measurement occasions. Results are discussed with reference to the measures` reliability.

(VAL) Validity and Reliability

Poster 112: SWITCH COSTS: CONSTRUCT VALIDITY AND 1-WEEK TEST-RETEST RELIABILITY

Veronik Sicard, Université de Montréal; Alexe Simard, Université de Montréal; Gabriel Lavoie, Université de Montréal; Robert Davis Moore, Arnold School of Health, University of South Carolina; Dave Ellemberg, Université de Montréal

The impact of concussions on an individual's cognitive functioning has become a growing health concern over the past several years; however, the search for sensitive tests persists. The task-switching paradigm is known to be sensitive to various medical conditions, including concussion. Accordingly, we developed two versions of the color-shape switch task. Three different costs are computed from the raw

scores: global switch cost, which is thought to be a measure of global cognitive control; local switch cost, which is believed to be a measure of cognitive flexibility; and working memory cost. The aim of this study was to evaluate psychometric characteristics of these costs. An ANOVA revealed a main effect of sex on local latency switch cost, with females exhibiting longer latencies (95.86 ± 11.23 ms) than males (61.01 ±10.94 ms), $p = 0.05$. No main effect of sex was observed on any other switch costs. Moreover, no main effect of experimenter or version of the task was observed. Local switch cost was significantly correlated with trails 4 and 5 of the Comprehensive Trail Making Test ($rs > .21$, $ps < .04$). No other significant correlation between costs and established neuropsychological tests was observed, indicating low convergent validity. The intraclass correlation coefficient estimates ranged from .23 to .77, suggesting low-to-moderate 1-week test-retest reliability. Therefore, the switch costs computed show low convergent validity and reliability. Clinicians and researchers may choose alternative ways to score the task-switching paradigms results, as proposed by Hughes and colleagues (2014).

(VAL) Validity and Reliability