University of Zurich[UZH]

Psychometric Society

IMPS 2017

# Abstracts: Posters

# Thursday, July 20, 2017

## Poster Session: 6:15 PM - 8:00 PM

**Poster 1: Modeling Circumplex Data with Latent Trait Models**
Ana Carla Crispim, University of Kent; Anna Brown, University of Kent

A circumplex is a two-dimensional orthogonal structure with points forming a circular pattern around the axes. Different models have been proposed in the past for circumplex data, however, results from these models do not seem to be reliable when these models are fitted to categorical data. The aim of the study was to verify the performance of latent trait models with categorical data with hypothesised circumplex structure. We simulated 68 categorical items (3-point scale) complying with a circumplex structure using function sim.circ from R psych package, varying the sample size (N=400, 600, 800). Latent trait models were fitted using Mplus. We verified the number of factors and the "wave" pattern of the correlation matrices, and assessed factor models for goodness of fit and the parameter recovery. Lastly, we assessed circumplex characteristics in an empirical dataset (N=422) of responses to 61 items (3-point scale) measuring core affect. Scree plots supported a two-factor structure for the simulated data. RMSEA showed that the approximate model fit was acceptable (<0.08), except for the model with N=400. Factor loadings followed a circular pattern around the two axes in the loadings plot and a "wave" pattern was identified in the heat maps. Similar circumplex characteristics were found in the empirical dataset. In conclusion, circumplex characteristics can be verified with latent trait models, given that proper parameter constraints are applied. Besides being familiar to many researchers, the factor analysis approach benefits from easy implementation in many software packages.

APP 1

**Poster 2: An IRT Analysis of the Growth Mindset Scale**
Brooke Midkiff, The University of North Carolina at Chapel Hill; Michelle Langer, American Institutes for Research; Cynthia Demetriou, The University of North Carolina at Chapel Hill; A.T. Panter, The University of North Carolina at Chapel Hill

Growth mindset has gained popularity in the fields of psychology and education, yet there is surprisingly little research on the psychometric properties of the Growth Mindset Scale (Dweck, 2008). This research presents an IRT analysis of the Growth Mindset Scale when used among college students in the United States. The sample come from responses to 5 surveys administered as part of The Finish Line Project – a U.S. Department of Education First in the World grant funded project that seeks to improve First Generation College Student (FGCS) access to, persistence in, and completion of postsecondary education through rigorous research into various programs and supports for FGCSs. The sample consists of 1260 individuals who completed the Growth Mindset Scale on one of 5 surveys. Of the 1260, 691 were FGCSs, 549 were non-FGCSs, 273 were currently enrolled, and 987 were recent graduates. IRT analysis is used to assess item fit, scale dimensionality, local dependence, and Differential Item Functioning (DIF). Because growth mindset is believed to be important for academic success among historically marginalized groups, it is important to know if the Growth Mindset Scale functions well among FGCSs. The reliability, validity, and item-functioning of the scale have not yet been examined among FGCSs. Lastly, though research exists on the impact of interventions on growth mindset, little psychometric research on the scale itself exists. The research presented here fills this gap in the literature by providing an IRT analysis of the Growth Mindset Scale.

APP 2

**Poster 4: Comparison of Generalized DINA Family Models with TIMSS 2007 Data**
Kazuhiro Yamaguchi, University of Tokyo; Kensuke Okada

A number of Cognitive Diagnostic Models (CDMs) have been developed over a period of time. The Generalized Deterministic Inputs Noisy And gate (G-DINA) model includes a broad range of other CDMs. The G-DINA model provides a unified view of CDMs. In addition, it makes it easy to examine the comparison between models. However, there are few studies that compare G-DINA family models with large achievement tests, especially the Trends in International Mathematics and Science Study (TIMSS) data. It is important to assess which model is best fitted to real data because it becomes a guide to application of CDMs. This study aimed to assess the fitness of G-DINA family models using TIMSS 2007 fourth grade mathematics data and explore the trend of a well fitted model. We employed the data of fourth-grade Japanese and American students. We chose booklet four and five as data, respectively. Attributes and Q-matrix were just the same as in the previous study by Lee, Park, and Taylan (2011), which applied the DINA model to TIMSS data. We compared the G-DINA model, DINA model, Deterministic Inputs Noisy input Or gate (DINO) model, Additive Cognitive Diagnostic Model (A-CDM), Linear Logistic Model (LLM), and Reduced Reparametrized Unified Model (R-RUM). The result showed that R-RUM was the best fitted Japanese and American data from perspective of AIC. R-RUM is a non-compensatory model like the DINA model. This result implied that TIMSS mathematics data requires multiple cognitive skills and their combinations for students to solve problems.

APP 4

**Poster 5: Exploring Test-Taker Scratchwork in Relation to Item and Person Characteristics**
Nazia Rahman, Law School Admission Council; Charles Lewis, Educational Testing Service; Chris Fox, Law School Admission Council

In paper-and-pencil testing, multiple-choice questions are typically presented in a test booklet and test takers record their final answers on a separate scannable answer sheet. Test takers are often permitted to use the space in their test booklets for scratchwork, which commonly consists of notes, diagrams, underlining of text, or other types of markings. Scratchwork and the relationship it may have with test-taker ability on a large-scale, high-stakes multiple-choice test is the focus of this research. Additionally, the relationship between item characteristics and the amount of scratchwork was also explored. If a relationship can be demonstrated between test-taker ability (as measured by the number of correct responses), item characteristics, and the amount of scratchwork on the test booklet, this information can potentially be used as secondary evidence to support questioning a test taker's score if cheating is suspected. As an added benefit, this exploration can also inform the development of a test-taker interface as the paper-and-pencil format transitions to other digital formats for test delivery. Three hundred randomly selected test booklets from three different test administrations were used. The 25th, 50th, and 75th percentiles of the score distribution were chosen to draw a random sample of about 100 test-taker registration numbers at each of these score levels. Preliminary analysis shows that the use of scratchwork generally increased with score level, both unconditionally and given a right response. However, at each score level, there were substantial differences across items and across test takers in terms of the use of scratchwork.

APP 5

**Poster 6: Using Item Response Theory to Inform De Novo Measure Development**
Sheree Schrager, Children's Hospital Los Angeles; Mary Rose Mamey, Children's Hospital Los Angeles; Jeremy Goldbach, University of Southern California

Although valid, reliable measurement is critical to explanatory research and intervention efforts, rigorous measure development remains a notable challenge. The predominant model for

understanding health disparities among Sexual Minority (e.g., lesbian, gay, and bisexual) Adolescents (SMA) is minority stress theory (Meyer, 2003), yet nearly all published studies of SMA rely on minority stress measures with poor psychometric properties. In a study of 346 diverse SMA ages 14-17, we demonstrate the utility of Item Response Theory (IRT) analysis to produce a new SMA minority stress measure with desirable properties including measurement invariance across multiple demographic subgroups. We applied IRT to a 72-item candidate item set. Discrimination and difficulty parameters and item characteristic curves were estimated overall, within each of 12 initially derived factors, and across demographic subgroups by age, gender, sexual identity, and race/ethnicity. Whole-scale difficulty values ranged from -1.210 to .572 and discrimination values ranged from 1.512 to 5.668 above the mean. Subscale analysis resulted in the removal of two items for excessive discrimination (443.20 and 16.17). The measure demonstrated configural and scalar invariance for gender and age; a three-item factor was excluded for demonstrating substantial differences by sexual identity and race/ethnicity. Three additional items were removed following reliability analysis. The final 64-item measure comprised 11 subscales and demonstrated excellent overall (α =.98), subscale (α range .75-.96), and test-retest reliabilities (whole-scale r>.99; subscale r range .89-.99). By using information from IRT models to guide item selection, we posit that construct measurement can be improved across all areas of psychological research.

APP 6

**Poster 7: Robustness of Country Rankings: A Curriculum-Adjusted Worst/Best-Case Sensitivity Test**

Stephan Daus, University of Oslo; Johan Braeken, University of oslo

A major concern in international large-scale assessments is fair comparisons of educational systems, irrespective of differences across the curricula. Despite that curriculum implementation is known to have a relatively strong association with student achievement, the influence of this indicator on country-level scores remains unknown. We investigated how robust the country comparisons and rankings in terms of student achievement would be when adjusting for differences in the degree of curriculum implementation across the participating countries. In particular, for the IEA-TIMSS 2011 science assessment in grade 8, we compared a total of 40 TIMSS countries, excluding benchmarking participants and countries with a 100% non-response rate on teacher-reported implemented curriculum items. As a worst-case / best-case sensitivity test, curriculum-implementation-adjusted country scores and rankings were computed based on either the minimally or maximally observed within-school implementation score for a country. To ensure comparability with the international reports, we followed in all our analyses the design-based statistical inference approach using plausible-value estimation of the science achievement and domain achievement measures in combination with total student sampling weight and replicate weights. In the process, we chart the country-specific curriculum implementation profiles across the TIMSS science domains and variability in curriculum implementation across the classrooms in the different countries. Results of this sensitivity test and the curriculum implementation profiles will be presented and discussed with particular attention to the countries with noticeable non-robust outcomes.

APP 7

**Poster 8: Gender Comparison of Achievement Growth Patterns: Different Subjects and Scores**

Xiaohong Gao, ACT, Inc.; Deborah J. Harris, ACT, Inc.; Wei Tao, ACT, Inc.

Standardized achievement assessments are universally the core of an accountability system. Research on gender differences indicate that girls often receive higher grades than boys regardless of school subjects although a myth from admission tests is that boys typically outperform girls on math and science. However, very few studies examine gender differences in growth patterns throughout the school years across different school subjects. This study investigates where "gender

gaps" may occur during the students' academic growth and whether the patterns are similar across school subjects based on individual and aggregated scores. The study uses rich multiple types of data from multiple content areas in a large-scale assessment program, using multiple item types and testing modes. In one data set, students in different grades took the same test, with content spanning multiple grade levels. The second data set is longitudinal, with the same students testing over multiple grades, each year with grade specific content. In the third data set, students took both multiple interim tests with a summative assessment. Since assessment scores are a sample of measures that students could obtain based on a sample of items scored by a sample of scoring methods, the current study also investigates whether the growth patterns are consistent under different scoring methods. Different types of scores derived from CTT and IRT are used in the analyses. Therefore, results of the study will not only shed light on whether there are gender difference in growth patterns but also whether the differences are consistent among different measures.

APP 8

**Poster 9: Effects of Private Tutoring on Students' Mathematics Achievement in China**
Xuran Wang, Beijing Normal University

Although private supplementary tutoring, so called shadow education, is expanding at an alarming rate in China, this phenomenon has escaped the attention of Chinese researchers, especially not prevalence in qualitative research. This study focus on private tutoring in four rapidly developing cities in China. It draws on secondary analysis of the 2015 Program for International Student Assessment (PISA) student and school survey databases. Hierarchical Linear Models (HLMs) was used to explore the relationships between private tutoring and mathematics achievement. Also, using Propensity Score Matching (PSM) to find the net effect of private tutoring to math achievement. In the four cities, mathematics private tutoring had positive effect to students' achievement. If students from different Socio Economic Status (SES) backgrounds could receive equal private tutoring opportunities, their achievement gap caused by SES could be narrowed. In addition, no gender difference was found between students who took part in private tutoring and those who did not. Finally, this paper highlighted the influence of individual and school factors on students' achievement, and reported on students' declared reasons for taking private tutoring. In conclusion, schools should provide necessary free private tutoring for students from lower SES backgrounds. And, large private tutoring enterprises should provide funds for students from lower SES families or for high-ability students.

APP 9

**Poster 10: Developing a Learning Progression of Fraction in China**
Yizhu Gao, Beijing Normal University; Tao Xin, Beijing Normal University

Learning progressions have been suggested as one vehicle to research the development of students' cognition towards core subject concepts. For this study, it applies the Rule Space Model, a kind of cognitive diagnostic model, to measure the learning progression of fraction for primary school students. This analysis firstly extracted eight cognitive attributes and their hierarchy model from the analysis of previous research and mathematics textbook. A hypothesized learning progression of four levels was built accordingly. Then, a cognitive diagnostic test for fraction addressing the attributes of understanding fractions, proper fractions, improper fractions, mixed fractions, greatest common factor, simplifying fraction, least common multiple, and applying the least common multiple was developed. Finally, the model was used to analyse a sample of 541 Chinese primary school students' observed item responses to identify their knowledge states and to validate and modify the hypothesized learning progression. The results showed that the test was of good psychometric quality by analyzing difficulties, discriminations, reliabilities, and validities. The students were classified into twenty-five attribute mastery patterns and learning paths were

provided for absolutely mastery of fraction. By applying the rule space model, the hypothesized learning progression was modified at each level.

APP 10

**Poster 11: A Bayesian Zero-Inflated Poisson Regression Analysis on School Security**
Meng Qiu, University of Maryland

While the amount of school-related deaths are beginning to decrease, incidents of theft and violence – including student violence against teachers – are on the rise in US's schools based the report, Indicators of School Crime and Safety: 2014, released by the National Center for Education Statistics (NCES). One of the key findings in this most recent report showed that "during the 2009–10 school year, 85 percent of public schools recorded that one or more crime incidents had taken place at school, amounting to an estimated 1.9 million crimes." "Our nation's schools should be safe havens for teaching and learning free of crime and violence," the report said. This demonstrates the necessity of school crime and security to be one of the major concerns of educators, policymakers, administrators, parents, and students. Analyzing and understanding the school policies and programs associated with school crime and security is an essential step in developing strategies to address the issues of school crime and violence. This is the motivation of this study. The purpose of this study was to investigate whether certain school attributes could facilitate reduction of school crime and improvement of school security using Bayesian Zero-Inflated Poisson regression (ZIP) model. A public national dataset from 2007-2008 School Survey on Crime and Safety (SSOCS) was analyzed.

BSI 1

**Poster 12: Testing the Emotional and Behavioral Problems of Adolescence Using a Bayesian Approach**
Namwook Koo, Korea Institute for Curriculum and Evaluation; Hyunchul Kim, Sungkyunkwan University; Ji Young Mun, Sungkyunkwan University; Eun Kyung Seo, Sungkyunkwan University; Kyung Lim Kwon, Sungkyunkwan University; Hyeyeon Park, Sungkyunkwan University; Meounggun Jo, Sungkyunkwan University

Emotional and behavioral problems in adolescence vary from simple and temporary maladjustment to serious psychological disorder (Ebata, Pertersen, & Conger, 1990; Kazdin, 1993). In South Korea, adolescent emotional and behavioral problems have emerged as a serious issue. This study investigates the change in emotional and behavioral problems among South Korean adolescents such as attention, aggression, somatic symptoms, social withdrawal and depression. We do not assume that all emotional and behavioral problems of South Korean students change over time. Thus, we use a Bayesian approach to support the null hypothesis of zero effect size. However, in Bayesian hypothesis testing exact computation of Bayes factors (BF) is difficult and new methods to compute BF and BF approximations have been developed. Thus, we compare different methods for computing BF such as a default prior distribution specification (Liang, Paulo, Molina, Clyde, & Berger, 2008; Rouder & Morey, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009) and BF approximations such as the Bayesian Information Criterion (BIC), Scaled Unit Information Prior Bayesian Information Criterion (SPBIC) and Information matrix-based Bayesian Information Criterion (IBIC) (Bollen, Ray, Zavisca, Jarden, 2012; Masson, 2011). We analyzed longitudinal data from the Korea Child Youth Panel Survey (KCYPS) where 2,280 students' emotional and behavioral problems were measured in two different time points of 2011 (grade 8) and 2013 (grade 10). We discuss computation of BF and Bayesian hypothesis testing from a practical standpoint.

BSI 2

**Poster 13: An Investigation of Bayesian Scoring Under Diagnostic Cognitive Models**
Yu-Lan Su, ACT, Inc.

Bayesian scoring approaches have been extensively studied under IRT models. However, under Diagnostic Cognitive Models (DCM), the existing literature is very limited in examining scoring consistency of mastery/non-mastery in student attribute profiles. This simulation study is devoted to investigating classification consistency of skill mastery for two Bayesian scoring methods in the realm of DCM. Bayesian EAP (Expected A Posteriori) and MAP (Maximum A P+AF104osteriori) are applied to estimate attribute profiles under DCM. This study employs the non-compensatory DINA model - a deterministic, inputs, noisy, "and" gate model, and the compensatory DINO model - a deterministic, inputs, noisy, "or" gate model. To be closer to the practical testing, the percentage of items measuring attributes in Q-matrix simulation is based on analyzing TIMSS Math test with National Council of Teachers of Mathematics Principles and Standards for School Mathematics (NCTM, 2000). Samples of 2,000 examinees are simulated for three item-by-attribute Q-matrices (i.e., 20-by-5, 40-by-5, and 40-by-8) for 50 replications for each model. The study also analyze scoring consistency (1) for the group as a whole, and (2) for high, middle and low ability groups. This study will report differences in profiles of master/non-mastery, classification consistency, and correlations of examinees' binary skill vectors for the EAP and MAP methods under DINA and DINO models separately. The results will help better understanding of the performance of two Bayesian scoring approaches on estimating students' attribute mastery, and hence to inform educational practitioners the importance of selecting scoring methods when applying DCM to student reports.

BSI 3

**Poster 14: Effect of Covariate Selection on Treatment Estimation Using Propensity Score Matching Techniques**
Heather D. Harris, James Madison University; Jeanne Horst, James Madison University

Propensity Score Matching (PSM) techniques provide a means by which researchers can control for selection bias through creation of a comparison group that is qualitatively similar to participants on key variables (i.e., "covariates;" Austin, 2011; Stuart, 2010). However, in order to adequately implement PSM techniques, a researcher must make a series of decisions, including choice of covariates (Steiner, Shadish, Cook, & Clark, 2010; Stuart, 2010; Stuart & Rubin, 2008). Although the PSM best practices literature suggests including covariates related to both the individual's decision to participate and the outcome (Brookhart et al., 2006), little is known about how different combinations of covariates affect the accuracy of inferences in educational research. In the current study, we evaluated the accuracy of treatment effect estimates when covariates unrelated to treatment outcome were included in the model. Data were simulated in R version 3.3.2 (R Core Team, 2013) and propensity score matching was used to create a comparison group for purposes of estimating the treatment effect. Both conservative (i.e., including only covariates related to both treatment and the outcome) and liberal approaches (i.e., including covariates unrelated to the outcome) were employed. Six conditions were simulated that varied the strength of relationships between covariates and the treatment outcome. The treatment effect was estimated across each of the six simulated conditions. Recommendations for the applied researcher and suggestions for future research are offered.

CAU 1

**Poster 15: Effects of CAT Scoring Methods on Comparability with Linear Forms**
Benjamin Andrews, ACT, Inc.

When examinees can take either a Computerized Adaptive Test (CAT) or a linear version of the same test, steps must be taken to ensure the comparability of the scores. One potential source of differences in the psychometric properties of the two types of tests is the way the tests are scored.

CATs are usually scored using an estimate of theta while linear tests are often scored based on the number of correct responses. This can sometimes lead to substantial differences in the properties of the scores. This research compared several different scoring methods including MLE estimates and EAP estimates using either the entire response string or the summed score. In addition, two number-correct scoring methods that use equating were used. The first is a method described by Stocking (1996) that uses IRT true score equating to adjust for differences in difficulty for items on the CAT. A similar method that uses IRT observed score equating to adjust for difficulty differences was also used. Results from simulations were evaluated using criteria outlined by Wang and Kolen (2001) for assessing the comparability between CAT and linear test scores. Specifically, first- and second-order equity for scale scores were compared among the different scoring methods. Other properties such as the similarity of the scale score distributions were also compared. Both passage-based tests and tests containing only discrete items were considered.

CBT 1

**Poster 16: Collaborative Problem Solving Skills Among Students with Learning Disabilities**
Chen-Huei Liao, National Taichung University of Education; Kai-Chih Pai, National Taichung University of Education; Ying-Hsien Li, Taichung Ssu Chang Li Elementary School; Bor-Chen Kuo, National Taichung University of Education

The present study aimed on developing a computerized assessment of Collaborative Problem Solving (CPS) and exploring the performance of CPS skills among children with Learning Disabilities (LD) and non-disabled children in elementary schools in Taiwan. In the proposed CPS assessment, students should communicate and co-work with one or more computer-based agents, and attempt to solve a problem. Three core collaboration abilities, establishing and maintaining shared understanding, taking appropriate actions to solve the problem, establishing and maintaining team organization; and four cognitive behavior processes, exploring and understanding, representing and formulating, planning and executing, monitoring and reflecting were investigated. CPS skills were assessed with four test units which were developed in the context of reading, math, and social studies. One-hundred and forty-four grade 5 and grade 6 students, and 46 fifth and sixth graders with learning disabilities were recruited from public schools in Taiwan. The results showed that the non-disabled children performed significant better than the LD group with the scores of overall CPS skills, the core abilities of collaboration, and cognitive behavior processes. A strong and significant relationship between the core abilities of collaboration and the cognitive behavior processes of problem solving to the overall collaborative problem solving skills was found.

CBT 2

**Poster 17: The Relationship Between Response Time and Probability of Answering**
Chih-Wei Yang, National Taichung University of Education; Bor-Chen Kuo, National Taichung University of Education; Mao-Hsiung Chen, National Taichung University of Education

The aim of this study is to discuss the relationship between response time and some item characters using real data. Multiplication of decimal tests were used to collect fifth grade students' responses and also response times of all items by a computer based test. Students have no time limits to answer the questions and are requested to write down their problem-solving processes for further analysis. The results of this study shows that the distribution of response time is similar to the log-normal distribution. The response times for the items differs according to the question type and the complexity of the correct problem-solving process. The correlations between response time and item parameters are also presented and discussed.

CBT 3

**Poster 19: Measuring Implicit Leadership Theories Using Online Card-Sorting of Facial Images**
Jing Rachel Ma, University of St. Andrews; David Ian Perrett, University of St. Andrews

Implicit leadership theories are internal knowledge structures of behaviours and traits that characterise an ideal leader. Traditional measurement of implicit leadership theories was criticised to elicit implicit schema using explicit methods, i.e. self-report of desired leadership characteristics. We demonstrated the possibility of using visual measurement to illustrate implicit leadership theories. We hypothesised that (a) people would be able to distinguish leaders from non-leaders on the basis of facial images (b) leadership preference would be context contingent, with different characteristics being desired in different leadership roles, and that (c) preferences would reflex rater's group differences. Using an online card-sorting interface, Caucasian and East Asian participants were asked to sort male and female Caucasian faces into one of the following categories: business leader, sports leader, moral leader, and not a leader. A "not sure" category was also included to allow uncertain categorisation. Face choices were compared with perceived personality ratings (e.g. masculinity, intelligence, attractiveness, health, maturity). We obtained the following results: (1) An individual's facial appearance predicted whether or not he/she will be considered as a leader. (2) Different sets of faces were chosen according to leadership contexts. (3) Contextual choice was driven by facial cues to different personality traits (e.g., perceived intelligence was important for business and moral leadership but not for sports leadership). (4) Cross-cultural agreement was found with subtle differences in the visual representations of leadership prototypes. In conclusion, facial appearance can be used to explore implicit leadership theory. Leadership preference patterns are context contingent and largely consistent across cultures.

CBT 4

**Poster 20: Differences between Computer Based and Paper & Pencil Testing**
Khurrem Jehangir, National Center for Assessment

Computer Based Testing (CBT) conveys certain advantages for the test administration bodies and this has been one of the reasons for the gradual phasing out of the traditional Paper and Pencil Tests (PPT) in many a setting around the globe. Nevertheless PPT administration is still retained by many entities and some of them also offer their examinees a choice between a PPT or CBT administration. While the shift from PPT to the CBT format maybe motivated by practical advantages for the exam administrator, what are the consequences for the examinees? This question is investigated using data on 5 similar test forms (containing the same items) from the CBT and PPT modes of administration of a University entrance exam administered to over half a million examinees. The test consisted of two parts, a verbal section and a quantitative section. A range of analyses were done for both sections separately. The comparative analyses included examining dimensionality and reliability. We also investigated the impact of a mode effect, the impact of DIF on test scores across the two modes, differences in violation of Local Independence across the 2 modes, differences in "constancy of theta"+AF48 (person-fit) at the beginning or end sections of the test for the two modes, and differences in item-cluster position effects across the two modes. Some systematic differences were observed between the CBT and PPT formats across the two modes for the 5 test forms. Furthermore, these differences did not always have the same direction for the verbal and quantitative test forms.

CBT 5

**Poster 21: Simulating a Bi-factor Computerized Adaptive Test to Measure Internalizing Psychopathology**

Matthew Sunderland, University of New South Wales

Comorbid mood and anxiety disorders affect a sizable proportion of the population and these disorders are associated with a substantial burden of disease, poor outcomes, and increased service utilization. Highly efficient assessments that better account for comorbidity between mood and anxiety disorders (e.g., internalizing) are therefore required to assess and monitor individuals who are most at risk of psychopathology in the community. The current study examined the efficiency and validity associated with a bifactor Computerized Adaptive Test (CAT) to measure broad and specific levels of internalizing psychopathology. The sample comprised 3,175 respondents who completed an online survey recruited through Facebook. Items from five banks (generalized anxiety, depression, obsessive compulsive disorder, panic disorder, social anxiety disorder) were jointly calibrated using a bifactor item response theory model. Simulations indicated that an adaptive algorithm could accurately ($r \geq 0.90$) estimate general internalizing and specific disorder scores using on average 44 items in comparison to the full 133 item bank (67% reduction in items). Scores on the CAT demonstrate convergent and divergent validity with previously validated short severity scales and could significantly differentiate cases of DSM-5 disorders. As such, the internalizing CAT (INT-CAT) efficiently measures comorbidity among the population in a more valid and empirically-informed manner than existing disorder-specific screening scales. The INT-CAT is available to administer to research participants and patients using an R program available from the lead author.

CBT 6

**Poster 22: Comparing Multidimensional to Unidimensional Computerized Adaptive Testing: The Bigger Picture**

Muirne Paap, University of Oslo; Sebastian Born, Friedrich Schiller University Jena; Johan Braeken, University of Oslo

Research has shown the benefits of taking into account the correlation among dimensions when estimating latent trait scores in Computerized Adaptive Tests (CATs). Multidimensional CATs (MCATs) could further improve measurement precision/decrease test length as compared to using separate unidimensional CATs for each domain, especially if domains are highly correlated. In this talk, we will provide the audience with a bird's eye view on CAT by presenting results from a systematic simulation setup used to establish under which conditions MCAT offers advantages over using separate Unidimensional CATs (UCATs). Variable-length CAT administrations (both MCATs and separate UCATs for each dimension) are simulated based on empirically derived item banks under scenarios typical for health and educational assessment. The simulation design factors are assessment area, item bank size, item type, and correlation among the three latent dimensions. The evaluation criteria are efficiency of the testing procedure (test length, stopping rule used) and quality of the latent trait estimation (bias and SE). Our results indicate that UCATs are already highly efficient in case of adequate targeting, but overall MCATs are even shorter. Results vary depending on appropriate targeting, item bank, and person scores. Bias is more pronounced when targeting is poor. MCAT has great potential when it comes to reducing test length and improving accuracy and precision of latent trait scores, both in health and educational measurement. We will discuss how the incremental value of MCAT depends on factors like adequate targeting, the size of the correlations, item bank size, and item parameters.

CBT 7

**Poster 23: Assessing Collaborative Problem Solving Performances of Taiwanese Students in Mathematics**

Shu Chuan Shih, National Taichung University of Education; Hung Chung Wu, National Taichung University of Education

Because Collaborative Problem Solving (CPS) is a necessary capability across educational settings and workplaces in the 21st century, this topic is increasingly appreciated in education. This study aims to develop an online collaborative problem solving assessment in mathematics based on the framework of PISA 2015 draft, and analyze the collaborative problem-solving performances of Taiwanese students in mathematics. There were 30 mathematical CPS multiple choice items and auto-scoring models included in the two test units (game of getting 25, Tower of Hanoi). Participants were 56486 ninth through tenth graders in Taiwan. According to the collected responses of the mathematical CPS items, the results of test analysis indicated that the test with reasonable reliability and acceptable item discrimination. Furthermore, students' CPS performance based on item response theory analysis and lag sequential analysis were also explored. The findings showed that higher level of performance in establishing and maintaining shared understanding and lower level of performance in taking appropriate action to solve the problem, female students' math CPS capability was significantly better than male students, and there were different CPS behavioral patterns among students of different ability level. These results could be used as references for future researches and indicators for teachers in math curriculum design.

CBT 8

**Poster 24: Relationship Between Item Exposure Rate and General Test Overlap Rate in CAT**

Shu-Ying Chen, National Chung Cheng University

Item exposure control is commonly adopted to reduce the threat of item sharing in computerized adaptive tests. Item exposure rate and general test overlap rate are two indices commonly considered in item exposure control. By considering these two indices, item exposure can be monitored at both item and test levels. The general test overlap rate is referred to the average proportion of common items between an examinee and a group of previous examinees. The general test overlap would be equivalent to the pairwise test overlap when the size of the group considered is one. When examinees collect test information from more than one of previous test takers, the general test overlap would be more appropriate than the pairwise test overlap to capture the larger scope of information sharing. Even though item exposure rate and general test overlap rate are defined differently, they are closely related and can not be controlled independently. When items are administered to all examinees, the item exposure rates would be as high as 1.0, and the general test overlap would not be low. On the other hand, when items are administered with strict control on item exposure rates, the general test overlap would not be high. To control these two indices effectively, the relationship between the two indices should be considered. The purpose of this study is to investigate the relationship between item exposure rate and general test overlap rate thoroughly, such that an effective procedure for controlling these two indices can be designed.

CBT 9

**Poster 25: Empirical Investigations of Position Effects on Pre-Equating for CAT**

Tony Thompson, ACT, Inc.

In transitioning a testing program from paper testing to Computerized Adaptive Testing (CAT), it is common for initial item parameter estimates of the item pool to come from calibrations based on the paper format. Using parameter estimates from a prior paper administration for pre-equating in CAT might raise concerns of item position context effects, especially if items are administered in fixed positions on paper. Item position effects may impact item parameter estimates and

ultimately examinee scores if items appear on the CAT in positions different than paper. The impact on parameter estimates is likely if the test is speeded to some degree, as items at the end of a speeded test would appear more difficult than if those same items appeared earlier in the test. This study examines a number of real data sources to document the degree of impact that position effects can have on parameter estimates and scores. These data sources include operational paper form data where items have shifted position as well as pilot study CAT data. The study also looks at ways to mitigate the effect on CAT, through either imposing position constraints on items in CAT or by modeling (e.g., Kang, 2014; Debeer & Janssen, 2013) how position change impacts item parameter values. The study examines both discrete item tests as well as passage-based tests. This research should be of interest to CAT practitioners as well as to researchers interested in item position effects.

CBT 10

**Poster 26: Models of Students' Age and Educational Crises**
Elena V. Leonova, Tsiolkovskiy Kaluga State University

The problem of students' disadaptation in the new educational environment during transition to a new level of education has been studied both theoretically and empirically. Our study focused on the problem of a pupil's agency during the periods of normative crises (age crises and crises of adaptation to the next levels of education). We hypothesized that students with a high level of agency (as a result of the age crisis) will use their psychological resources to solve adaptation crisis problems. Whereas students with a low level of agency will not use their resources in full: this category of students has problems in learning, behavior, negative emotional state despite their intellectual capacity. 383 students took part in the study. By using the k-means method of cluster analysis, each category of students was divided into clusters with similar agency indicators. Adaptation criteria (information, behavioral and affective) were defined in accordance with B.F. Lomov's systemic approach. Regression models of adaptation and disadaptation of students were developed. General, age and individual psychological factors of students' disadaptation (first-graders, fifth-graders, tenth-graders and freshmen) were defined. It was proved that low level of agency indicators (personal qualities, motives, values) – are the psychological factors of disadaptation for all categories of students. Age-specific disadaptation problems of students were determined. It is proven that pupils and students, who adapted successfully, have high levels of life meaningfulness and the desire for self-actualization. Overall, the results suggest that agency is an important factor of adaptation crises' overcoming.

CCC 1

**Poster 27: Model-Based Clustering for Tensor Data**
Kohei Uno, Osaka University

We propose Model-Based Clustering (MBC) for tensor data. Tensor methods have attracted much attention for a long time in psychometrics and chemometrics. These days, tensor methods are increasingly gaining popularity not only in psychometrics but also in machine learning. Generally, multivariate methods are applied to two-way data because these methods are based on the probability distribution of random vectors. Our proposed method is based on the multilinear-variate normal distribution instead of the normal distribution. The multilinear-variate normal distribution is an extension of the matrix-variate normal distribution. Our proposed procedure is assessed in a simulation study and illustrated with real data examples.

CCC 2

**Poster 28: Clusters of Response Behavior in Time-Limited Multiple-Choice Intelligence Testing**
Natalie Borter, University of Bern; Olivier Pahud, University of Bern; Thomas Rammsayer, University of Bern

Raw scores on time-limited multiple-choice intelligence tests are determined by two error types: incorrect responses and errors of omission. The aim of the present study was to search for different clusters of response behavior based on these two error types. For this purpose, 69 participants completed a time-limited multiple-choice intelligence test. By means of a k-means cluster analysis, three subgroups were identified: (1) a few-error cluster characterized by few errors of omission and few incorrect responses (n = 31), (2) an error-of-omission cluster with many errors of omission but few incorrect responses (n = 9), and (3) an incorrect-response cluster with many incorrect responses but few errors of omission (n = 29). The few-error-cluster showed the highest mental ability, answered the easy items quickly, and committed few errors. The error-of-omission cluster showed lower mental ability than the few-error-cluster, spent much time on answering the easiest items and committed many errors of omission. The incorrect-response cluster showed similar mental ability to the error-of-omission cluster, but exhibited fast responses to the easiest items and showed many incorrect responses. These differential patterns of response behavior clearly indicated that response time cannot be considered a generally valid indicator of an individual's level of mental ability. Moreover, our findings challenge the common, popular notion of a negative relation between response latencies for easy items and performance on time-limited multiple-choice intelligence tests.

CCC 3

**Poster 29: Elimination versus Negative Marking in Multiple-Choice Tests: An Empirical Comparison**
Rianne Janssen, KU Leuven; Qian Wu, KU Leuven; Jef Vanderoost, KU Leuven; Tinne De Laet, KU Leuven

In high-stakes exams, multiple-choice questions are often dichotomously scored on correctness, but with a penalty for choosing a distractor. This negative marking discourages students from guessing, but in case they do not know the correct response, they are not able to show their partial knowledge. In elimination scoring (Coombs, Milholland & Womer, 1956) students indicate the smallest subset of alternatives of which they are confident to contain the correct response by eliminating all incorrect responses. This scoring method gives credit to partial knowledge (Ben-Simon, Budescu & Nevo, 1997) and the scoring rule of Arnolds and Arnolds (1970) controls the expected gain due to guessing. The present study compares both scoring methods for the same multiple-choice questions using a within-subjects design for a low-stakes exam in the history of psychology and a between-subjects design for a high-stakes exam in medicine. Apart from comparing the obtained scores across students and across items, also a score group analysis is performed comparing for each item the relative response frequencies of the possible answer patterns for low- and high-ability students. One of the findings is that elimination instructions result in students showing less "no knowledge"+AF74 but also slightly less "full knowledge" than the traditional instructions (cf. Chang, Lin & Lin, 2007). It is also shown that for some items low-ability students are performing about equally well as high-ability students in recognizing the correct response alternative, but they are less able to distinguish it from the distractors.

CTT 1

**Poster 30: Item Purification Versus Adjustements for Multiple Comparisons in DIF Detection**
Adéla Drabinová, Charles University & The Czech Academy of Sciences; Patrícia Martinková, The Czech Academy of Sciences; David Magis, University of Liege

Most classical Differential Item Functioning (DIF) detection methods rely on the basic principle of testing for DIF one item after each other, which can have very strong impact on the identification of DIF in terms of power, rejection and Type I Error. In an extensive simulation study we compared six different scenarios of controlling Type I Error, including item purification and multiple comparison adjustement methods (Holm and Benjamini-Hochberg). Three DIF detection methods were selected: the Mantel-Haenszel test, the logistic regression and Lord's chi-square test based on 2PL IRT model. Examining empirical rates as well as using beta regression models, early results suggest that the effect of used correction methods is various in different DIF detection methods. Generally, adjustement procedures caused decrease of both rejection and power rates. On the contrary, purification led to slight increase of power when rejection rates were not really affected. Our results can be summarized in suggestions under which circumstances item purification and when multiple comparison correction methods (or even their combination) should be used.

DIF 1

**Poster 31: Identifying One Credible Referent Variable for Measurement Invariance Testing: A MIMIC-Interaction Modeling**
Cheng-Hsien Li, National Sun Yat-sen University; Kwanghee Jung, ACT, Inc.

The fulfillment of measurement invariance is considered as a prerequisite for meaningfully proceeding with substantive cross-group comparisons. However, two potential limitations related to model identification purposes in the multiple-group CFA approach, unfortunately, have received little attention among methodologists and substantive researchers: (1) the specification of a referent variable (i.e., standardization) in the test of factor loading invariance and (2) the lack of a statistical test for intercept invariance (see, e.g., Millsap, 2011; Raykov, Marcoulides, & Li, 2012; Rensvold & Cheung, 2001). A Multiple-Indicator Multiple-Cause (MIMIC) model with linear and moderated effects (i.e., a MIMIC-interaction modeling approach: Woods & Grimm, 2011) to detect uniform and non-uniform measurement biases in tandem is proposed here to identify a single credible referent variable, aiming at solving the two aforementioned issues. A Monte Carlo simulation study was carried out to determine the effects of different configurations of number of noninvariant variables, location of noninvariance, magnitude of noninvarince, magnitude of group differences in factor means and variances, and sample size in a unidimensional measurement model. Data generation and analysis were performed with Mplus. The true positive rate was used to assess the performance of MIMIC-interaction modeling in correctly identifying one credible referent variable from truly invariant variables in the population. Results showed that the proposed method performed well across almost every condition, except when equal to or more than 40% of observed variables were manipulated as non-invariant across groups, suggesting that the proposed procedure of using a MIMC-interaction model is practically recommendable.

DIF 2

**Poster 32: An Exploratory Strategy to Identify Sources of Differential Item Functioning**
Chi-Chen Chen, National Sun Yat-sen University; Chung-Ping Cheng, National Cheng Kung University; Ching-Lin Shih, National Sun Yat-sen University

The items being deemed as exhibiting Differential Item Functioning (DIF) should be carefully revised by item writers. If more information can be provided to item writers, such as DIF status and sources of DIF, the revision of items could be done more efficiently and precisely. All the current quantitative methods used to identify the sources of DIF items assumes that all possible sources should be known in advance and usually being collected together with the item response data.

However, this is not always the case in reality. An exploratory strategy that can used to explore the source of DIF, which might be more flexible for real test conditions, is proposed in this study. The performance in identifying sources of DIF of this strategy, combined with the MIMIC method, is investigated through a series of simulation studies. The results supported that given a set of DIF-free items that defined the main dimension can be correctly identified and the proposed exploratory MIMIC method can correctly recover the number of sources of DIF, the items that belong to each source of DIF can be recovered almost perfectly. This strategy can be applied to different DIF assessment methods and its performance is investigated in another simulation. It is found the hit rates of correctly identified the number of sources of DIF increased mainly as the true positive of DIF assessments method increased. The results and findings of this study are further discussed.

DIF 3

**Poster 33: Multilevel Measurement Invariance: A Monte Carlo Simulation**
Elizabeth Svoboda, University of Nebraska - Lincoln

Testing multilevel measurement invariance has gained attention in the social, behavioral, and education sciences. The pervasiveness of testing in modern society renders the utilization of a method for establishing comparative judgments essential. Comparative judgments can be considered across groups and across hierarchical levels of analysis simultaneously. Decisions must be made to allocate resources to schools across states in the United States (across groups) as well as to teachers' classrooms within schools (across hierarchical levels of analysis). Measured data are often multilevel in nature. Multilevel measurement invariance considers comparative judgments between groups and within levels of analysis. Multilevel measurement invariance tests whether constructs have equivalent properties and relations to one another at the individual as well as cluster levels of analysis by simultaneously constraining levels of invariance to be equivalent across clusters (Kim, Kwok, & Yoon, 2012). The current study examined the effect of measurement invariance testing while correctly accounting for nested data. The Type I error rate and statistical power of testing invariance at the cluster-level and individual-level simultaneously were investigated. The design decisions (intraclass correlation, number of clusters, cluster size, and frequency of noninvariance) that affect Type I error rate and statistical power in multilevel data were studied. Multiple-group multilevel confirmatory factor analysis was used to determine the effect of correctly modeling the dependence between observations in multilevel data. Monte Carlo simulations indicated as the intraclass correlation increased, Type I error rate increased contingent upon number of clusters more than cluster size.

DIF 4

**Poster 34: Comparing Different Methods for Detecting Differential Testlet Functioning**
Guan-Yu Chen, Beijing Normal University; Ping Chen, Beijing Normal University

A testlet is a bundle of items that share a common stimulus (Wang & Wilson, 2005a), and more and more educational and psychological assessments have started to adopt testlet design. Although testlets can reduce the burden of examinees and improve the efficiency of tests, it violates the local independent assumption that is basic for Item Response Theory (IRT). Thus, applying the IRT-based Differential Item Functioning (DIF) analysis to the testlet items may produce bias. Some researchers have proposed several methods to investigate the Differential Testlet Functioning (DTLF), however the relative merits of the various methods have not been evaluated in the literature. We perform such an evaluation, using simulated and empirical data. This research analyzes the DTLF under the framework proposed by Beretvas and Walker (2012). Because a testlet item DIF consists of testlet-based DTLF and item-based DIF, we will use the exiting methods such as Rasch testlet model (Wang & Wilson, 2005b), bifactor multidimensional IRT model (Fukuhara & Kamata, 2011), and two-level testlet response model (Beretvas & Walker, 2012), to deal with the testlet items. These methods will

also be compared with the SIBTEST method (Douglas et al., 1996) in a simulation study under different scenarios: no DIF or DTLF, both DIF and DTLF, and cancellation (DIF and DTLF are inconsistent in direction). The two-level testlet response model is expected to yield the best results, and then it will be used to analyze the real data from the National Basic Education Assessment for Chinese 4th and 8th grade students.

DIF 5

**Poster 35: Controlling Extreme Response Style in DIF Assessment Using a MIMIC Approach**
Hui-Fang Chen, City University of Hong Kong; Kuan-Yu Jin, The Education University of Hong Kong; Wen-Chung Wang, The Education University of Hong Kong

Extreme Response Style (ERS, a tendency to endorse extreme response categories, regardless of item contents) is prevalent in self-report measures and leads to underestimated or overestimated latent ability. Traditional Differential Item Functioning (DIF) assessments are vulnerable to ERS, because participants are matched on a biased matching variable. This study proposed a new model under the framework of Multiple Indicators Multiple Causes (MIMIC), which includes the variance of the observed scores for each respondent as the indicator of ERS to eliminate the impact of ERS in DIF detection. A series of simulations were conducted to evaluate the performance of the modified method against the conventional MIMIC method by manipulating four factors: constant or variant item slopes, test length (10 or 20 items), impact (0 or 1), and the difference in ERS between the two groups (none, small to moderate, and large). The responses of 500 participants in the focal and reference groups were simulated respectively, and false positive rates were evaluated. Results showed that ERS caused inflated false positive rates in the conventional MIMIC approach, and the situation became worse as the difference in ERS between two groups and the test length increased. The inflations were controlled at the nominal level when the new MIMIC model was implemented. Additionally, neither the performance of the traditional nor that of the new MIMIC model was influenced by item slopes or impact.

DIF 6

**Poster 36: Information Criterion as Effect Sizes for Differential Item Functioning**
Teck Kiang Tan, Institute for Adult Learning

Various effect sizes for DIF have been recommended in the literature such as Cohen's standardized mean score, Mantel-Haenszel common log-odds ratio, R2 based effect size and Raju's area method. This paper proposes a new approach in measuring effect sizes using information criterion. Change in information criterion, say ΔAIC, between two logistic regression models is used as effect size. A positive value of ΔAIC indicates that the logistic model for a DIF regression model improves over a non-DIF regression specification, whereas a negative value shows that the added DIF covariate has reduced the information content, and a zero value indicate no differences in the information content. A large positive value of a ΔAIC thus indicates model with DIF specification is a better model choice whereas zero and small negative value indicates that there is no DIF. Data from a sample of 3,725 grade 10 students from a Mathematics assessment were used for this study. Outliers identified from a ΔAIC distribution are classified as large effect size otherwise small effect size. Bootstrapping generates the distribution of ΔAIC. As non-Gaussian is expected, three robust methods for identifying outliers suggested by Wilcox (2005) were used. They are the Interquartile range, Carling's modification (Carling, 2000), and the MAD-Median rule. Carling's modification shows more conservative results compared to the other two methods. However, it is less conservative than other effect size methods such as R2 based effect size, Dorans, Schmitt and Bleistein (DSB), and the ETS Delta scale (ETS) under various DIF methods.

DIF 7

**Poster 37: On the Model Mis-Specification and Identifiability Constraints for MC-FA**
Yu-Wei Chang, Feng Chia University

The Multiple-group Categorical Factor Analysis (MC-FA) Model is one of the common tools for understanding group differences for ordered categorical data or polytomous items. Due to the over-parametrization in the model, we need identifiability constraints to make the model just-identified. However, we argue that the identifiability constraints in such multiple-group models are mis-leading. It seems that some constraints put further restrictions on the model. In the current study, we characterize the parameter subspace resulting in model mis-specification, or, in the terminology of item response theory, failing to achieve concurrent calibration. In addition, we clarify that some set of constraints never result in model mis-specification. The investigation is important for practical applications since the parameter subspace leading to model mis-specification under MC-FA is quite different from our intuition.

DIF 8

**Poster 38: On Selection Bias in Repeated Cross-Sectional Surveys for Estimating Trends**
Melanie Wall, Columbia University; Aaron Sarvet, Research Foundation for Mental Hygiene, Inc.

Repeated cross-sectional national surveys are commonly used to monitor population trends for outcomes of interest in epidemiology, education, and demography. Typically, a large sample is collected and weights are used to obtain population representative estimates of survey responses. A similar sampling strategy is repeated across years to allow tests of trends across time. For example in epidemiology, trends in past-year substance use in the U.S. have been assessed using the National Survey on Drug Use and Health (NSDUH) from 2002-2013 and also the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) in 2002 and 2013. Prior methodological development has shown that careful consideration of age, period, and birth cohort effects is important for elucidating the different causal factors influencing trends. Selection bias (i.e. systematic differences in the relationship between the sample and the population) that differs across time is a factor that also may influence estimated trends. In the recent context of substantial and increasing survey non-response, it is especially important to consider this important source of bias for valid trend estimation. In this work we present a method for identifying the magnitude of differential selection bias in repeated cross-sectional surveys by comparing estimates of benchmark variables that should remain fixed across survey years within birth cohorts in the absence of differential selection processes. We demonstrate how to statistically control estimates of trends in order to lessen the deleterious effect of differential selection bias. The method is shown using NSDUH and NESARC data examining trends in substance use.

ECM 1

**Poster 39: Combining Factors from Different Factor Analyses**
Aniko Lovik, KU Leuven; Vahid Nassiri, KU Leuven; Geert Verbeke, KU Leuven & Hasselt University; Geert Molenberghs, Hasselt University & KU Leuven

While factor analysis is one of the most often used techniques in psychometrics, comparing or combining solutions from different factor analyses can be cumbersome even though combining factors is necessary in several situations. For example, when applying multiple imputation (to account for incompleteness) or in case of multilevel data (where a simple solution is the application of multiple outputation) often tens or hundreds of results have to be combined into one final solution. While different solutions have been in use, we propose a simple and easy to implement solution to match factors from different analyses based on factor congruence. A modified Tucker's congruence coefficient is used to match factors. To demonstrate this method, the Big Five Inventory (BFI) data collected under the auspices of the Divorce in Flanders study was analysed, combining multiple imputation with multiple outputation and factor analysis. The data were

collected in 2008 and the validated Dutch language version of the BFI was administered among a battery of tests, with the aim to study the phenomenon of divorce in families. This multilevel sample consists of 7533 individuals coming from 4460 families with about 10% of incomplete observations.

FAC 1

**Poster 40: Empirical versus Arbitrary Cut-Off Points in Exploratory Bi-Factor Target Rrotation**
Eduardo Garcia-Garzon, Universidad Autónoma de Madrid; Francisco J. Abad, Universidad Autónoma de Madrid; Luis E. Garrido, Universidad Iberoamericana

Empirically based target rotations, in which the target is based on an initial standard rotated factor solution, have been recently introduced within exploratory bi-factor analysis. Traditionally, the target matrix is determined on the basis of an arbitrary cut-off point (i.e., .30) for distinguishing substantive and trivial factor loadings, what may result in target misspecification and impaired performance (e.g., when combining a high cut-off point with the presence of weak factors). Based on Promin (Lorenzo-Seva, 1999), this study presents a bi-factor rotation that estimates factor-specific, befitting cut-off points: The Schmid-Leiman (SL) with iterative target rotation based on empirical factor loading cut-offs (SLiE). Due to the combination of its iterative nature and the empirical based cut-offs, this new algorithm was expected to outperform alternative SL bi-factor rotations, which either applied non-iterative target rotation with empirical cut-offs (SLtE) or iterative target rotations with arbitrary cut-off points (SLi). A simulation study that manipulated eight variables was carried out in order to evaluate the performance of SLiE, SLi (with arbitrary cut-off points ranging from .15 to .35), SLtE and bi-geomin. The results showed that SLiE recovered the bi-factor structures across the majority of conditions. On the other hand, the performance of SLi was dependent upon the cut-point selection: with cut-off points as high as .20, it performed similarly to SLiE. However, when conservative cut-off points (.30 or higher) were chosen, its performance was severely diminished. In general, SLiE provided the biggest improvements when weak factors were present, and all SL-based rotations were superior to bi-geomin.

FAC 2

**Poster 41: On the Bias in Eigenvalues of Sample Covariance Matrix**
Kentaro Hayashi, University of Hawaii at Manoa; Ke-Hai Yuan, University of Notre Dame; Lu Liang, Florida International University

Principal Component Analysis (PCA) is a multivariate statistical technique frequently employed in research in behavioral and social sciences. Often, PCA is used for approximating Exploratory Factor Analysis (EFA) because the former is easier to implement. In the context of PCA, Lawley (1956) showed that if eigenvalues of population covariance matrix are distinct, then the mean of the eigenvalues of the sample covariance matrix have a bias term of order 1/N. When the number of variables p is negligible relative to the sample size N, the sample eigenvalues are asymptotically unbiased. However, when p is large, the bias term is no longer negligible. We show that under some regulatory conditions, the order of the bias term is p/N.

FAC 3

**Poster 42: How Students' Perspectives on Social Problems Impact Their National Identity**
Weilin Jin, Peking University; Yujia Liu, Beijing Normal University

National identity is the social psychological basis for a regime to gain legitimacy and is important for a nation to exist, sustain and develop. With the processes of globalization, modernization and urbanization, crisis of national identity has become a critical global issue. In existing theories, government performance, history and culture are main influencing factors of national identity.

However, in past studies, effects of different kinds of factors on national identity should be discussed in detailed classifications. At the same time, current research has not paid enough attention towards East Asian and Southeast Asian countries. In this study, we use data from Asian Student Survey (2013) on 4113 Asian undergraduate students, and find out that student perspectives on social problems can serve as a measurement of governance performance as well as social order and stability. We use factor analysis to divide the different kinds of social problems into 5 groups: problems of basic need, lack of labor, conflicts, immorality and injustice, overpopulation. We find that students' different perspectives on social problems influence their national identity differently. A strong national identity is more common when concerns regarding basic needs and lack of work increase, while it is less common when concerns regarding immorality increases. Perspectives on conflicts and population problems do not pose a significant threat to national identity.

FAC 4

**Poster 43: Model Fit Performance of CDMs for Small Sample Sizes**
Hueying Tzou, National Taiwan University; Ya-Hui Yang, National Taiwan University

Cognitive Diagnosis Models (CDMs) are psychometric models for assessing examinees' mastery and nonmastery of skills or attributes in order to provide more fine-grained learning information. In particular, lectures are becoming given in small classes due to the impact of the low birth rate in Taiwan. Teachers could pay attention to smaller class sizes and provide more specific learning feedback to students by applying CDMs to analyze students' testing responses. However, practitioners often face the difficulty of choosing an appropriate model from a large number of CDMs and how to construct a correct Q-matrix. In addition, misspecified Q-matrices have produced bad model fit, thus related approaches have been developed to detect and to correct for misspecification in the Q-matrix. For example, Chiu (2013) developed the minimum residual sum of squares and de la Torre & Chiu (2016) developed ζ2 index, both approaches work well in Q-matrix correction (the Q-matrix recovery rate is higher than 90%). However, the performance of ζ2 index in smaller sample sizes and higher Q-matrix misspecification rates have not been not studied yet. We will explore three issues in this study: a) the performance of the ζ2 index in conditions with small sample sizes and high Q-matrix misspecification; b) the correct selection rates (selecting the correct model and Q-matrix combination ) of several model fit indices (AIC, BIC, r, l) for small sample sizes; and c) the correct selection rates of model fit indices in conjunction with a modified Q-matrix.

FCM 1

**Poster 44: Ordinal and Disordinal Within-Subject GxE Interactions under Measurement Error**
Ralph C. A. Rippe, Leiden University

Individuals who vary in their responsivity to the environment do so according to a) diat+AF72hesis stress (DTS) (Monroe & Simons, 1991; Zuckerman, 1989), where some individuals are more susceptible to negative consequences of adverse experiences than others, or b) differential susceptibility (DS) (Bakermans-Kranenburg & Van IJzendoorn, 2007; Belsky & Pluess, 2009), where some individuals are more susceptible to both negative and positive consequences than others. Susceptibility markers include genetic, temperamental, and physiological factors (Belsky et al., 2007; Ellis et al. 2011). Effects of measurement error on both predictor and outcome side are evident in terms of attenuated effects and larger standard errors (Hutcheon, Chiolero & Hanley, 2010). Rippe (IMPS 2016) investigated the effect of measurement error when testing a ordinal interaction (DTS) against a disordinal one (DS) (following Widaman et al., 2012). Proposed corrections for linear error exist (Bisbe et al., 2006; Huang, Wang & Cox, 2005), but corrections for tests of (disordinal against ordinal) within*between subject interaction are not provided. Therefore, we first study interactions in a within-subject design, to indicate under which conditions such corrections are empirically required. Through FPR and power calculations, we assess the effect of

varying sample size and increasing error for predictor, moderator and outcome, on the ability to accurately test competing interaction models. Results show that, for varying effect (F=.10 to F=0.40) and sample sizes (n=15 to n = 250) (based on Faul et al., 2013), small but realistic amounts of error already obscure detection of interactions and the difference between them.

FCM 2

**Poster 45: Investigation of Performance of S-X2 for Bi-Factor and Second-Order Model**
Yu-lim Kang, Yonsei University; Gue min Lee, Yonsei University

The benefits of IRT applications such as item banking, computerized testing, and test equating might not be attained if the fit between IRT model and data is not satisfactory. Numerous statistical procedures have been developed to evaluate IRT models and goodness-of-fit. Two commonly reported statistics are Pearson statistic and likelihood ratio. However, these statistics are known to be sensitive to test length and sample size and cannot avoid sparseness problem in the underlying contingency table when calculated. To resolve these issues, Orlando and Thissen (2000, 2003) suggested S-X2 statistics conditioned on a summed score instead of latent trait. Orlando and Thissen's procedure has found to adequately control Type I error and power for dichotomous item response (Suarez-Falcon, 2003; Stone & Zhang, 2008). Also, Kang and Chen (2008) extended S-X2 to polytomous IRT models and reported it performed properly compared to traditional item fit indices. Even though, achievement and aptitude tests consist of item bundles or testlets are widely used, testing goodness of fit for IRT models containing testlet-based data remained relatively underdeveloped. The purpose of this study is to investigate the Type I error rate control and power of the item fit proposed by Orlando and Thissen (2000, 2003) under 2PL UIRT, Bi-factor and second-order model. For simulation, manipulated factors include sample size, test length and testlet effect.

FCM 3

**Poster 46: Effect of Difficulty Distribution Between Testlets on Linking Testlet-Based Tests**
Bu Wenjuan, Beijing Normal University; Wen Hongbo, Beijing Normal University

Current methods for linking tests using traditional Item Response Theory (IRT) methods assume local independence. However, when tests are constructed using testlets, one concern is the violation of the local item independence assumption. The Testlet Response Theory (TRT) model and bi-factor model can be used to accommodate local item dependence. Many studies suggest that the TRT model and bi-factor model perform better than the traditional IRT model when linking testlet-based tests, but these results may not remain valid when the difficulty distribution between testlets is different. This study intends to explore the effect of the difficulty distribution between testlets on model selection for linking testlet-based tests using real data. Results of the study indicate that when the difficulty distribution is consistent between testlets, the traditional two-parameter logistic IRT model performs almost as well as or even better than TRT model and bi-factor model. However, when the difficulty distribution is discrepant between testlets, the traditional two-parameter logistic IRT model generates larger equating errors compared to the TRT model and bi-factor model. Furthermore, linking separate calibration yields smaller equating errors than concurrent calibration.

IRT 1

**Poster 47: Practical Implications of IRT Model Misfit for ADHD Clinical Context**
Daniela R. Crișan, University of Groningen; Rob B. K. Wanders, University of Groningen; Don van Ravenzwaaij, University of Groningen; Rob R. Meijer, University of Groningen; Catharina Hartman, University of Groningen

Item Response Theory (IRT) is a modern measurement framework which is gaining popularity in mental health research. Its applicability is very broad, ranging from scale construction and assessment through estimating person trait scores that describe a person's standing on the trait of interest. There are many advantages to using IRT models over classical models, but there is also a price to pay: Like any other statistical model, IRT models impose some restrictions on the data, and model parameter estimates can only be trusted if the model assumptions are approximately met. However, IRT models and their assumptions represent ideals about data that are difficult to meet in practice. Thus, an important question is: How severe are the consequences of using estimates (e.g., person trait levels) derived from a poorly-fitting model on the practical decisions (e.g., person ranking, classifications into diagnostic groups, predictions)? We focused on research on Attention Deficit/Hyperactivity Disorder (ADHD). The psychometric properties of ADHD scales have mainly been investigated by means of Classical Test Theory and Confirmatory Factor Analysis, and person scoring has been mainly based on sum scores. Little has been done thus far with respect to using IRT models to understand and improve ADHD assessment. Important research topics include assessing: Scale quality, person scoring and diagnosis, how well trait level estimates relate to external variables, and how would diagnosis of ADHD and predictions change by dropping potentially poorly-functioning items. We will present some exploratory results that hopefully shed some light on these problems.

IRT 2

**Poster 48: Optimal Designs for Pairwise Comparison in Education**
Elise Crompvoets, Tilburg University & Cito; Anton Béguin, Cito; Klaas Sijtsma, Tilburg University

Pairwise comparison is a scaling method for preference data. Data of comparison of pairs of objects on a trait are analyzed to create a rank order of the objects on this trait. Procedures to collect this type of preference data have been known for a long time, but have only recently received attention in education (Pollitt, 2012). In education, administering a complete design of paired comparisons (e.g., of writing assignments) is often unfeasible for raters (e.g., teachers) due to the large number of comparisons to be executed. To reduce their number, incomplete designs are used in which comparisons are selected based on heuristic adaptive algorithms. Because current algorithms may be suboptimal, we developed an adaptive algorithm that maximizes the information from the comparisons. In a simulation study we investigated the impact of incomplete designs on rank order accuracy and on the uncertainty of the parameters. We compared a complete design with both a random incomplete design and the newly-developed adaptive incomplete design using various conditions. First, we varied the number of comparisons in the two incomplete designs as a percentage of the number of comparisons in the complete design. In addition, we varied the number of objects and consequently the number of comparisons because we expected that the influence of the percentage of comparisons on the uncertainty of the parameters varies with the number of objects. The commonly-used Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959) was used to analyze the data. Results from the simulation study are presented.

IRT 3

**Poster 49: The Radicals of Abstract Reasoning Assessments: An Explanatory Item Response Modelling Approach**
Fredrik Helland-Riise, University of Oslo; Johan Braeken, University of Oslo

Abstract reasoning assessments are widely considered the gold standard of intelligence testing. Although Raven Progressive Matrices (RPM) are probably the most prominent example, there are many more less studied and less developed abstract reasoning assessments used in practice. To improve these assessments and generalize RPM study results, we need a better understanding of what factors of the item design drive item response behavior. In modern test design such factors are called radicals, item properties that when modified lead to for instance a change in item difficulty; this in contrast to incidentals, item properties that are mere cosmetic changes not affecting psychometric item characteristics. Here, we use an explanatory item response modelling approach to examine the effects of potential radical item properties derived from a cognitive lab and from an artificial intelligent item solver to provide validity evidence for an non-Raven-like abstract reasoning assessment. We discuss how this validity evidence can be a first step in a more systematic redesign of the assessment opening up opportunities for automatic item generation and computerized adaptive testing.

IRT 4

**Poster 50: Performing Model Comparison via a Bayesian Procedure**
Henghsiu Tsai, Academia Sinica; Ya-Hui Su, National Chung Cheng University

Although many model comparison methods have been proposed in previous studies, selecting an appropriate model is still an important issue for Item Response Theory (IRT). In this study, we propose a Bayesian procedure to perform model comparison. First, a Bayesian procedure is used to fit a 3PL model for all multiple-choice items in a test. Then, we test whether a 2PL model would be appropriate for any specific item. Specifically, we claim a 2PL model is an appropriate model, when the credible interval of the guessing parameter of an item includes zero. The Bayesian procedure will be investigated through simulations. A real data set from a Taiwan national test will be analyzed as well.

IRT 5

**Poster 51: Searching Minimal Conditions for Bayesian IRT Equating in Non-Optimal Matrix-Sampled Design**
Hyun-Woo Nam, SoonChunHyang University

The purpose of this study is to find hierarchical prior information for stable estimation and equating for IRT parameters even when research data is composed of matrix-sampled anchor items with multidimensional mixed-format. The objective is to search for the optimum Bayesian conditions for specifing hyper-priors hierarchically when estimating and equating IRT parameters from test data, which is difficult to solve by the existing maximum likelihood estimation method. Simulation data are generated based on the 2006-2007 years of the 10th grade English test of the Korean National Assessment of Educational Achievement, assuming the 2-parameter logistic model for dichotomous items and the graded response model for polytomous items. When simulating data, 3 factors are manipulated, i.e. number of respondents, ability levels, and dimensionality. The number of respondents of each block is varied by 1,000, 500, 100, and 50. The ability level is set to the average level, below average level, above average level. The test dimensionality is supposed to be uni-dimensional, 2-dimensional, or 3-dimentional. IRT parameters are estimated and equated by varying the level of precision or hyper-priors of variance in a hierarchical manner. IRT parameters are estimated and equated using WinBUGS. The results are evaluated by calculating the RMSD (Root Mean Square Deviation) and bias. The theoretical and practical meaning of hierarchical hyper-priors compared to independent or fixed hyper-priors is discussed.

IRT 6

**Poster 52: Bayesian Estimation of the Multidimensional Nominal Response Model**
Javier Revuelta, Universidad Autónoma de Madrid; Carmen Ximénez, Universidad Autónoma de Madrid

This poster+AF39 introduces Bayesian estimation and evaluation procedures for the multidimensional nominal response model. The utility of this model is to perform a nominal factor analysis of items that consist of a finite number of unordered response categories. The key aspect of the model, in comparison to the traditional factorial model, is that there is a factor loading for each response category on the latent traits, instead of having one factor loadings associated to the items. The extended parameterization of the multidimensional nominal response model requires large samples for estimation. When sample size is of a moderate or small size, some of these parameters may be weakly empirically identifiable and the estimation algorithm may run into difficulties. In this investigation we propose a Bayesian MCMC inferential algorithm to estimate the number of latent traits underlying the multidimensional nominal response model. Two Bayesian approaches to model evaluation were compared: discrepancy statistics (DIC, WAICC and LOO) that provide an indication of the relative merit of different models, and the standardized generalized discrepancy measure that requires resampling data and is computationally more involved. A simulation study was conducted to compare these two approaches, and the results show that the standardized generalized discrepancy measure can be used to reliably estimate the dimensionality of the model whereas the discrepancy statistics are questionable. The poster also includes an example with real data in the context of learning styles, in which the model is used to conduct an exploratory factor analysis of nominal data.

IRT 7

**Poster 53: Person Response Function with Response Style Parameters**
Krzysztof Fronczyk, University of Warsaw; Ewa Witkowska, The Maria Grzegorzewska University

The presentation shows a model describing the parametric Person Response F+AF50unction (PRF), which can include some characteristics of the response styles in the questionnaires. This is possibly due to the decomposition of the person parameter theta (threshold) as the sum of all weighted response styles characteristics included into the model in a similar way as in the case of latent regression models. The model is applied to dichotomous responses. The Bayesian approach was used to estimate model parameters. The model was tested using simulated data and then applied to analyze data of six dimensions measured by Formal Characteristics of Behaviour—Temperament Inventory (FCB—TI) by Zawadzki and Strelau (1993) – Activity, Emotional Reactivity, Perseverance, Endurance, Briskness and Sensory Sensitivity. Social desirability, acquiescence and individual variability of the responses were included as response style characteristics. It was concluded that comparing to 1-PL model person response function, the described above model better fits the data for a majority of the FCB—TI scales. The presented model is appropriate in all situations where response sets of individual persons are important.

IRT 8

**Poster 54: Effects of Overlapping Constructs of Composite Score CSEM**
Kyoungwon Bishop, University of Wisconsin-Madison; H. Gary Cook, University of Wisconsin-Madison

This study investigates effects of overlapping constructs across domains when Conditional Standard Errors of Measurement (CSEM) of composite scores are computed. The Standards for Educational and Psychological Testing (AERA, APA, and NCME, 2014) recommend CSEM in units of each reported score. Composite scores are often formed through a combination of weighted, summed test scores over two or more content areas. When content areas share similar constructs

such as a language test, the assumption of independence of errors must be questioned. Kolen and Lee (2010) found that CSEM estimates were larger using the unidimensional than the multidimensional model, and concluded that the unidimensional model led to overestimation of the CSEM. When correlated errors across domains are added, a composite CSEM might be inflated. Kolen, Wang, and Lee (2012) estimated composite CSEM using a multivariate latent ability distribution but still based on the assumption of independence of errors and added errors across domains. The WIDA consotium's English language assessment (ACCESS for ELL 2.0) consists of four domain tests: Listening, Reading, Speaking and Writing which are intended to represent one underlying construct of language ability. The four domain scores are highly correlated. This study approaches CSEM computation at the item level instead of scores of domains using ACCESS for ELL 2.0 data. We present the effects of different covariances and weightings on CSEM values with mixed item types, for unidimensional and multidimentional latent ability distribitions.

IRT 9

**Poster 55: Scale Alignment in Multidimensional Item Response Models**

Leah Feuerstahler, University of California, Berkeley; Mark Wilson, University of California, Berkeley

Multidimensional item response models allow for simultaneous measurement on multiple latent dimensions. After fitting a multidimensional model, we may want to compare examinee scores across dimensions. The interpretation of these comparisons is confounded with how the item response model is identified. Traditionally, multidimensional item response models are identified by centering the mean ability on each dimension at zero. In this scenario, comparisons across dimensions are relative to the mean ability on each dimension. However, there is usually no a priori reason to believe that mean examinee ability is equal across dimensions. Indeed, different dimensions may represent skill sets that differ in terms of the sophistication and rate of acquisition of the measured abilities (see e.g., Osborne et al., 2016). For instance, consider a math test that contains both algebra and geometry items. Although ability in algebra and geometry both reflect a general ability in math, students may have greater abilities in algebra than in geometry. In this case, we may want to compare dimension scores directly, instead of relative to the mean score on each dimension. In order to directly compare ability scores and item parameters across correlated dimensions, the dimensions need to be realigned. This study presents several techniques that linearly transform multidimensional item parameters such that units on each dimension can be compared in terms of absolute ability (Schwartz & Ayers, 2011; Schwartz, Ayers, & Wilson, 2014). The performance of these techniques is compared on real and simulated data.

IRT 10

**Poster 56: Using MIRT to Understand Situational Judgement Tests with Polytomous Responses**

Li Guan, University of Georgia & Development Dimensions International; Donald E. Lustenberger, Development Dimensions International

Researchers have argued that Situational Judgment Tests (SJTs), commonly used by organizations for high-stakes selection, are multidimensional in nature (e.g., Lievens & Sackett, 2007). Only recently, however, have attempts been made to model responses to SJT items using multidimensional IRT (e.g., Whelpley, 2014). Current research has yet to explore this modeling issue for SJT that employ polytomous responses; however, modeling the multidimensional nature of SJTs could help purify the measurement of underlying psychological constructs and potentially improve their criterion-related validities. Analyses were conducted on a set of 15 SJT items completed by 1000 incumbents from 4 organizations. Age of participants ranged from 22 to 73 ($M = 47.52$, $SD = 8.79$); 737 participants were males and 263 were females. An exploratory factor analysis was conducted prior to the MIRT analyses to detect the possible dimensionality of the SJT, and four dimensions were detected. Four Multidimensional Graded-Response Models (MGRM) were run. AIC, BIC and -2 Loglikelihood were calculated. No significant discrepancies were found among these

four models. Validities of CTT, 1-factor, and 2-factor model scores with a criterion measure were obtained and ranged from .32 to .35. The correlation between CTT and IRT 1-factor model scores was .86 ($p < .001$). Additional validities and correlations will be presented at the conference. Regression weights and relative weights are suggested to be an appropriate approach to identify SJT models with dichotomous responses (Whelpley, 2014); the investigation of polytomous responses will be presented at the conference as well.

IRT 11

**Poster 57: Comparison of Some Speededness Detection Strategies on IRT Parameter Estimation**
Lu Wang, ACT Inc.; Robert Ankenmann

It is acknowledged that the presence of speededness may lead to poor parameter estimation of unidimensional IRT models. Various methods were developed to reduce the speededness effects on the application of these IRT models, among which removing speeded examinees seemed as a feasible option in practice. Previous studies have shown that when response time information was available, setting response time threshold methods could be employed as effective ways in identifying speeded examinees (Wise & Ma, 2012; Wang, Liu & Gao, 2014). However, little information was found about to what degree these methods could improve parameter estimation. Therefore, the purpose of this study is to explore and compare the effects of three detection methods (see Wang, Liu & Gao, 2014, for more details) on parameter estimation across varying test conditions. Two sample sizes (N = 1,000 and 5,000) and three test lengths (L = 30, 50, and 70) are implemented. Under each test condition, speeded and nonspeeded examinees' responses, response times and true abilities are generated by manipulating van der Linden's hierarchical model (van der Linden, 2010), which incorporates response time as collateral information in a unidimensional IRT model. The detection methods are then applied to identify speeded examinees under each test condition. After excluding those speeded examinees, the classification consistency and parameter estimates under traditional 2PL and 3PL models are examined, respectively. The results of this study are expected to provide some applicable options for practitioners when speededness is a concern but unidimensional IRT models are employed in practice.

IRT 12

**Poster 58: Ideal Point IRT: Comparing the GGUM and Normal PDF Models**
Megan R. Lowery, University of Georgia; Alexandra Harris, University of Georgia; Nathan T. Carter, University of Georgia

The Generalized Graded Unfolding Model (GGUM; Roberts, Donoghue, & Laughlin, 2000) and the normal Probability Density Function (PDF) model (Maydeu-Oliveras, Hernandez, & McDonald, 2006) represent two distinct conceptualizations of the ideal point response process within the context of Item Response Theory (IRT). The two models differ in terms of complexity, the number of response options allowable, and the possibility for estimation of multidimensional solutions in popular software programs. The GGUM is a more complex model that allows for either dichotomous or polytomous responses, and likely better reflects the complex reality of ideal point response processes. However, the software currently available for the GGUM only allows for unidimensional models. In contrast, the normal PDF model is a simpler model that can only be applied to dichotomous responses but allows for convenient multidimensional estimation in its current software implementation in the 'mirt' package in R. Whereas the GGUM has seen considerable application in the personality, attitude, and vocational interest literatures, the normal PDF model has seen little attention. The capability for multidimensional models and the availability of absolute fit statistics familiar to broad audiences (e.g., TLI, RMSEA) in the 'mirt' package is likely to be of interest to many researchers. Thus, the current Monte Carlo simulation study asks whether the normal PDF model is capable of faithfully modeling the complexities inherent in the highly-

parameterized GGUM. Moreover, we address whether the practice of dichotomizing truly polytomous ideal point data for the purpose of applying the normal PDF model is appropriate.

IRT 13

**Poster 59: Comparing Threshold Models for Response Styles: A Simulation Study**
Mirka Henninger, University of Mannheim; Thorsten Meiser, University of Mannheim

There exists a variety of IRT approaches for rating data aiming to correct for response styles, such as the tendency towards extreme categories (ERS) or midscale responses (MRS). The multidimensional response style model and the random threshold model account for response styles by representing them as additional latent traits. While the former represents theoretically predefined response styles and allows for covariation between trait and response styles, the latter accounts for response tendencies heuristically through person-specific threshold shifts and assumes unrelated traits and threshold shifts. In a simulation study, we examined how different response style models react to varying covariation between trait and response styles. We generated data from population models with different covariance structures among the traits and response styles and analyzed the simulated data with the Partial Credit Model (PCM), the multidimensional model as well as the random threshold model. Overall, the level of covariation between trait and response styles seems to play a minor role for parameter recovery and model fit. Furthermore, results suggest that differences between models in recovering item and trait parameters were negligible, with the PCM performing slightly worse than the other models. However, the multidimensional model outperformed the random threshold model in recovering true response style parameters. Concerning information criteria, the two response style models showed a better fit than the PCM, however, the random threshold model needed the most iterations to converge. Implications concerning model choice for psychometric research and applications are discussed.

IRT 14

**Poster 60: The Uses of "Metric" in Modern Test Theory**
Núria Duran Adroher, Swiss Paraplegic Research; Birgit Prodinger, Swiss Paraplegic Research; Carolina Saskia Fellinghauer, Swiss Paraplegic Research; Alan Tennant, Swiss Paraplegic Research

In the two paradigms of Modern Test Theory (Rasch Measurement Theory [RMT] and Item Response Theory [IRT]) the term "metric" is frequently used, but what it actually means is hardly specified. This study aims to examine the use and implied meaning of "metric" in RMT and IRT. A systematic search and review with the terms (Rasch OR "item response" OR IRT) AND ("metric" OR "interval scale" OR "conjoint measurement" OR "fundamental measurement") was performed in the health and social sciences literature. The search was restricted to 2001-2005 and 2011-2015. The paradigm of each article, as well as its type (theoretical, methodological, teaching, or application) were identified. Descriptive statistics were computed and paragraphs elucidating the meaning of "metric" were extracted. Out of the 2639 eligible articles, 1337 were included. 944 were dealing with RMT, 293 with IRT, and 100 with both. 68% were application papers. In RMT, 413 (44%) articles mentioned "metric" and 716 (76%) "interval scale", being 275 (94%) and 37 (13%) in IRT. "Metric" was used as a synonym of "scale" (e.g., "common metric" equivalent to "common scale") and of "interval" (i.e., "metric scale"). "Interval metric" and "ordinal metric" were also mentioned. "Metric" was sometimes determined by specifying a mean and standard deviation (i.e., T-metric). In conclusion this study shows that (1) there is a wide range of uses of "metric" implying different meanings, e.g. "scale" or "interval property"; (2) "metric" is used more in IRT than in RMT. No consensus appears as to its mathematical properties.

IRT 15

**Poster 61: Investigating Free-Baseline DIF Detection Method of the Multidimensional Forced-Choice IRT Model**

Philseok Lee, South Dakota State University; Seang-Hwane Joo, University of South Florida; Stephen Stark, University of South Florida

Psychometric developments in Multidimensional Forced Choice (MFC) modeling have enabled the widespread use of noncognitive MFC measures. Although modern methods of test construction and scoring have overcome ipsativity problems historically associated with MFC formats, there has very little research on methods for examining measurement invariance, which is a preeminent concern in operational testing environments. As Brown and Maydeu-Olivares (2014) suggested, Differential Item Functioning (DIF) analysis is needed to provide test constructors with more statistical information at the item-level, support multinational testing efforts, and permit meaningful cross-cultural comparisons. This study proposed and evaluated a "free-baseline" MFC DIF method involving three-alternative MFC items based on the Thurstonian IRT model (TIRT; Brown & Maydeu-Olivares, 2011). Our simulation study investigated the effects of sample size (500 and 1000), magnitude of DIF (0.3 and 0.6), type of DIF (loading, threshold, and both), number of DIF items (1, 2, and 3), and number of statements having DIF within an MFC item (1, 2, 3 statements). DIF items were identified using a Wald-test comparing free-baseline (augmented) and compact models. Results indicated that this method was highly effective in detecting MFC DIF items, with power approaching 1.0 in the large sample size and large magnitude of DIF conditions. Power increased overall as the number of statements having DIF within an MFC item increased, and power was higher when DIF caused by differences in loading, rather than threshold, parameters. This presentation will review the TIRT model, the proposed free-baseline TIRT DIF method, and the simulation results and practical implications.

IRT 16

**Poster 62: A Comparison of Four Approaches for Testing the Rasch Model**

Rudolf Debelak, University of Zurich

In the psychometric literature, various statistical tests have been proposed to assess the overall goodness-of-fit of the Rasch model. Among these tests, first-order tests, which are based on the comparison of expected and observed scores for individual items, can be discerned from second-order tests, which are based on the comparison of expected and observed scores for item pairs. A well-known first-order test is based on the LR statistic (Andersen, 1973), whereas the M2 statistic (Maydeu-Olivares & Joe, 2005) is an example for a test statistic for a second-order test. The aim of this study is to compare these first- and second-order parametric tests with nonparametric first- and second-order tests for the Rasch model (Ponocny, 2001). In a variety of simulation studies, all model tests are evaluated with respect to their Type I error and their power against various model violations of the Rasch model. In these simulation studies, the Rasch model, a 2PL model, a multidimensional 1PL model, a 1PL model with a guessing parameter and a 1PL model with a learning parameter were used as data-generating models. The results indicate that the evaluated tests vary substantially with regard to their power against the various alternative models. The consequences for practical data analysis are discussed.

IRT 17

**Poster 63: Polytomous IRT Model Misspecification and Metric Implications**

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

Practitioners often select among polytomous IRT models based largely on personal preference, as it is usually difficult to distinguish models using goodness-of-fit criteria. However, model selection can have substantial effects on the underlying metric. This paper illustrates how misspecification may impact the underlying latent proficiency scale. We generate data from a 2PL-Sequential

Response Model (2pl-SRM; Mellenbergh, 1995) and subsequently fit the Generalized Partial Credit Model (GPCM; Muraki, 1992) and the Graded Response Model (GRM; Samejima, 1969) under the following conditions: increasing, decreasing, or constant item discrimination and difficulty parameters across steps. We compare the estimated and true examinee's abilities. The findings show substantial shrinkage or expansion at the low or high ends of metric in the presence of model misspecification, indicating that the interval-level properties can be substantially degraded when the polytomous model is misspecified.

IRT 18

**Poster 64: Multidimensional IRT and the Measurement of Student Growth**
Sora Lee, University of Wisconsin- Madison; Daniel M. Bolt, University of Wisconsin-Madison

IRT studies of growth often focus on creation of a vertical scale assuming a common latent ability across grades. Fundamental questions relate to the consistency of the latent ability across grade levels (e.g., is the same ability measured in Grade 3 the same as in Grade 4?), and item characteristics related to sensitivity to growth (e.g., are highly discriminating items within grade necessarily highly discriminating indicators of growth?). We consider item response data collected on two different tests for common students across two successive years, and two types of confirmatory two-dimensional MIRT models, referred to as prospective and retrospective change models. Each model distinguishes traits corresponding to (1) a baseline ability, defined as latent ability at either the higher or lower of grade levels, and (2) a change dimension, defined as either the forward change from the lower to higher (prospective model) or the backward change from the higher to lower (retrospective model) grade level. By allowing different loadings on both dimensions, we can account for the possibility that items may function differently as indicators of change than of baseline ability. From real data study, model comparison criteria confirm discrimination differences under both models. This suggests a change in the nature of ability measured over time, and item's discrimination within grade is often not a good indicator for measure of growth. Simulation analyses examine parameter recovery under several conditions: (1) the mean and variance of the growth dimension; (2) the magnitude of discrimination differences across traits; and (3) sample size.

IRT 19

**Poster 65: Multidimensional Graded Response Model Parameter Recovery: Under Nonnormal Trait Distribution**
Ummugul Bezirhan, Columbia University; Ngalula Fleurant, Columbia University

The Multidimensional Graded Response Model (MGRM) is widely used within psychology and education to analyze items that are designed to measure more than one latent trait and when responses are classified into ordered categories. However, it should be noted that the estimation software used for this model assumes normal latent distribution within its algorithm (MML/EM) which contrasts what is normally found in educational and psychological assessment. The latent trait within these fields is unlikely to be normally distributed. Furthermore, despite research that investigates how nonnormal latent distribution affects parameter recovery in other Item Response Theory models, less is known concerning the performance of MGRM with nonnormal latent trait distributions. To address this, we simulated data from different nonnormal latent distributions crossed with two different scaled intercorrelations while implementing two estimation methods: Marginal Maximum Likelihood (MML) and Markov Chain Monte Carlo (MCMC). Correlation, bias, and Root-Mean-Square Error (RMSE) is used to assess the quality of the parameter recovery. The results of this study will assist researchers in making informed decisions when using the MGRM in their data collection and analysis.

IRT 20

**Poster 66: Assessing Equating Transformations in IRT Observed-Score and Kernel Equating Frameworks**
Waldir Leôncio Netto, University of Padua; Marie Wiberg, Umeå University

The application of test equating methods to enable the comparison of test scores across different administrations implies the adoption of a set of statistical assumptions. The current literature acknowledges that selecting the best framework can be very challenging, since each alternative has its own set of assumptions, references and ideal outcomes. Once we view equating transformations as statistical estimators, though, such comparisons become feasible. This paper compares the statistical and computational properties of three equating methods, namely Item Response Theory Observed-Score Equating (IRTOSE), Kernel Equating and Kernel IRTOSE, under two different data-generating scenarios. Numerical and real data applications for those six combinations yield promising results. Moreover, they provoke a reflection about which of the best-performing combinations is the most adequate overall. Some practical suggestions as well as propositions for further research are also included.

IRT 21

**Poster 67: Application of MCMC in IRT With Flexible Latent Trait Distributions**
Xue Zhang, Northeast Normal University; Chun Wang, University of Minnesota; David J. Weiss, University of Minnesota; Jian Tao, Northeast Normal University

Normality of latent traits is a routine assumption made when estimating item parameters for Item Response Theory (IRT) models, but it might be unrealistic with some datasets. The purpose of this research was to present a new Markov chain Monte Carlo (MCMC) method for ordinal items with flexible latent trait distributions (i.e., normal, skewed, and bimodal). Specifically, the Davidian Curve (DC) was used to approximate the distribution of latent traits. We evaluated the performance of the proposed MCMC algorithm with DCs via a simulation study. The manipulated factors included the number of response categories, sample size, distribution of the latent trait, and the order of the Davidian curve. Average Root Mean Square Error (RMSE) and average bias were used to judge the accuracy of estimation, the Hanna-Quinn (HQ) criterion was used to choose the best DC order, and Integrated Squared Error (ISE) was used to evaluate the similarity between the true and estimated latent distribution. Preliminary results indicated that the MCMC algorithm with DCs could fit a normal and bimodal distribution well and a skewed distribution reasonably well and the method provided good estimates of item parameters.

IRT 22

**Poster 68: The Effects of Calibration Methods on IRT Vertical Scaling**
Yeonbok Park, Yonsei University; Guemin Lee, Yonsei University

In some cases, tests are developed for the purpose of comparing student growth across grades and/or ages. Those tests differ in difficulty, but are intended to measure similar constructs. Vertical scaling should be conducted to place student performance scores on the same scale. Vertical scaling+AF99 based on IRT is a complex procedure that requires various decisions. Previous studies indicated calibration method would be a critical factor in vertical scaling, but there would be still controversy about which calibration method adequately measures student's growth. The primary purpose of this study is to investigate the effect of different calibration methods (separate, concurrent and fixed item calibration) on growth pattern, scale variability and effect size of IRT vertical scales. Real data analysis and several simulations under various conditions were implemented with two types of sample size, number of common items and three different proficiency score distributions. The specific research objectives of this study are as follows: To compare growth pattern, scale variability and effect size of vertical scales constructed by three calibration methods (separate, concurrent and fixed item calibration) under different conditions.

Based upon the results of real data analyses, growth patterns were quite similar among three calibration methods but scale variability was somewhat different among three calibration methods.

IRT 23

**Poster 69: Comparing Scoring Methods for IRT-Based Multidimensional Forced-Choice Assessments**

Yin Lin, University of Kent &CEB; Anna Brown, University of Kent; Mathijs Affourtit, CEB

IRT-based assessments typically use one of three scoring algorithms: the Maximum Likelihood (ML) estimator (Birnbaum, 1958, 1968), the Maximum a Posteriori (MAP) estimator (Samejima, 1969), and the Expected a Posteriori (EAP) estimator (Bock & Mislevy, 1982). While the ML estimator suffers from outward bias and unboundedness, the two Bayesian estimators are biased towards the mean of the chosen prior. A fourth scoring method has been proposed that promises to reduce estimation bias and also address the unboundedness problem of ML. The Weighted Likelihood (WL) estimator was developed by Warm (1989) and extended to the multidimensional case by Wang (2014). Both authors conducted simulation studies demonstrating the superior performance of the WL estimator, specifically in the case of the unidimensional 3PL model (Warm, 1989), the unidimensional generalized partial credit model (Wang & Wang, 2001), and the multidimensional 2PL model (Wang, 2014). While these results are highly relevant, the behaviour of WL is yet to be seen in Multidimensional Forced-Choice (MFC) assessments using the Thurstonian IRT model (Brown, 2011, 2013). The MFC response format asks respondents to provide ranking to a block of two or more items, with each item representing a different latent trait. The rankings are then decomposed into dichotomous pairwise comparisons for modelling, with structural local dependencies when there are multiple pairwise comparisons within the same block. This simulation study compares the performances of the four estimators in such assessments under a number of conditions, including varying test lengths, latent trait relationships, and the proportion of negative items.

IRT 24

**Poster 70: The STARTS Model: A Reconsideration and an Expanded Analytical Framework**

Enrico Perinelli, Sapienza University of Rome; Guido Alessandri, Sapienza University of Rome

Despite the growing body of research regarding the study and the application of Latent State-Trait (LST) models, few studies focused their attention to Kenny and Zautra's (1995) Stable Trait-Autoregressive Trait-State (ST-ART-S) framework. One of the advantages of STARTS model is the possibility to take into account the autoregressive component of the construct (often underestimated in LST models; Prenoveau, 2016), and to be able to deal also with constructs assessed with a single indicator. In the present contribution, drawing on previous work by Anusic et al. (2012), we propose a rational procedure to implement, step by step, alternative versions of the STARTS model, namely the ART-S and the ST-S, both nested within the larger STARTS model. Our procedure offers the researcher the ability (a) to deal with convergence problems when fitting the full model (offering possible explanations about their source), (b) to deal with single indicator constructs, (c) to represent different theories about the longitudinal stability of psychological constructs, (d) to represent and to solve, in an unique analytical framework, different substantive research questions regarding the relationship between stable and transient factors of two constructs. To explain the advantage of our analytical framework in detail, we present an applied example in which univariate and multivariate models are used to address the issue of whether the Name Letter Test (NLT) is a valid measure of state-like self-esteem. Our results demonstrated the validity of this improved procedure, and its usefulness for the applied researcher.

LDA 1

**Poster 72: DLVQ Method for Assessing Missing Weights in CFA Model**
Liang-Ting Tsai, National Taiwan Ocean University; Chin-Kuo Wu, National Taiwan Ocean University; Cheng-Chieh Chang, National Taiwan Ocean University

This study proposes Dynamic Learning Vector Quantization networks (DLVQ) weighting adjustment to overcome the shortcomings of raking and Classical Raking Ratio E+AF55stimator (CRRE) methods. Two numerical simulation studies are conducted to examine the accuracy of confirmatory factor analysis parameters. The design factors for the Monte Carlo study were missing proportions (10% and 20%), sampling sizes (500, 1000, 1500, and 2000), and variation of groups(R=0.1, 0.2, 0.3, and 0.4), and three weighting adjustment methods (raking, CRRE, and DLVQ). Based on the evidence provided by these two simulation studies, Raking and CRRE both achieved near 95% coverage rates under the condition in which all of cell observations were positive. However, when one of cell observations was zero, the trends of coverage rates decreased when the variations of groups and sampling sizes increased. Correspondingly, the disadvantages of Raking and CRRE could be improved considerably by the proposed DLVQ. DLVQ had near 95% coverage rates among the all experimental conditions under both conditions when all of the cell observations were positive and when one of the cell observations was zero.

MIS 1

**Poster 73: On Call or in Emergency. Control the Vertigo of Missing data**
María Paula Fernández García, University of Oviedo; Guillermo Vallejo Seco, University of Oviedo; Pablo Esteban Livacic-Rojas, University of Santiago; Ellián Tuero Herrero, University of Oviedo

Missing data is the data for+AF59 variables whose values could not be recorded although this was planned (Carpenter & Kenward, 2008). It is undeniable that missing data are always a serious problem. With incomplete data, the amount of information is less, inferences are likely to be biased, and data are not representative of the population that one aims to study. This will be more severe when the loss of data is greater and when the mechanism that causes it is more aggressive. So, if we have this problem we can only do one thing, consider different techniques available to fix ithe problem, and choose the best one. In this "emergency situation", any solution, even the best possible solution, may not be good. There are situations where the loss of test power is immaterial in relation to the selection bias that may have been caused. For this reason, it is preferable for the researcher to be "on call" from the same time in that it begins to take shape the research so that the loss of data is present only in an anecdotal manner. We present a panoramic view of the different ways in which missing data show up in two scenarios of extreme vulnerability, clinical trials, and research in educational settings, and we offer guidelines that can help "control the vertigo" that can produce missing data.

MIS 2

**Poster 74: Use of the Intraclass Correlation Coefficient in Multilevel Modelling**
Hsiu-Ting Yu, National Chengchi University

The ratio of the between-group variance to the total variance is called the Intraclass Correlation Coefficient (ICC). It tells you the proportion of the total variance in dependent variable that is accounted for by the clustering. However, ICC can also be interpreted as the correlation among observations within the same cluster. In practical applications, the ICC has been commonly used to determine whether or not a multilevel modelling approach is necessary. However, the usages of ICC in the context of multilevel modelling are much more versatile than just an index that quantifying the degree of resemblance between individuals in the same group. This presentation will review and explore both the theoretical concepts of ICC and various practical applications of ICC in the context of multilevel modelling. This presentation will first review the history and the developments of the ICC. The meanings of ICC used as an index of dependency and an index of

"variance-explained-for" are explained and compared. Factors that affecting the degree of ICC are also discussed and demonstrated using simulated data. One useful application of ICC as a model selection tool in multilevel models is then presented and illustrated. Finally, how ICC should be considered in designing multilevel studies and its relationship to the sample size and the power analysis are also included in this presentation.

MLM 1

**Poster 75: Determinants of Reading Achievement Level in PISA 2015**
Jung-A Han, Yonsei University & Korea Institute for Curriculum and Evaluation; Yulim Kang, Yonsei University

The OECD Programme for International Student Assessment (PISA) measures knowledge and skills of 15-year-old students every three years. The first PISA study took place in 2000 and PISA 2015 is the sixth in OECD's series of international assessments. Korea has been involved since PISA 2000. The PISA 2015 results of Korean students were at a high level among participating countries. However, the average scores of Korean students for three main domains in PISA 2015 decreased as compared to the results of PISA 2012. Especially, the proportion at the low level (below level 1) sharply increased in reading. Also, the proportion at the high level (above level 5) slightly decreased. Accordingly, it was concluded that the overall score decrease was caused by the increase at the low level. It is necessary to distinguish characteristics of students according to each achievement level. Therefore, the purpose of this study is to explore the factors affecting the reading achievement level of Korean students. To accomplish this, a Binomial Logistic Multi-level model was applied to estimate both student and school level effects. The independent variables involved in the analysis of the high level and the low level are the same.

MLM 2

**Poster 76: Using Dominance Analysis to Determine Predictor Importance in Multilevel Models**
Razia Azen, University of Wisconsin - Milwaukee; Luciana Cancado, University of Wisconsin - Milwaukee

Dominance Analysis (DA) was developed to determine the relative importance of the predictors in general linear models (Budescu, 1993; Azen & Budescu, 2003). The DA procedure examines all possible subset models formed from a set of predictors and compares the incremental fit obtained when each predictor is added to each subset model. One predictor is said to dominate another if it produces a larger incremental fit over all subset models or, more weakly, on average across models. Further, the consistency of the dominance relationships can be evaluated by examining the rate of reproducibility of the sample dominance results over many bootstrap samples. An extension of the procedure for multilevel models has been proposed (Luo & Azen, 2013) but has not been evaluated systematically. In this study we conducted a simulation to evaluate the performance of DA with multilevel models. To evaluate how well predictor importance is captured by DA under known conditions, we manipulated the correlations among the predictors and their relationship to the criterion variable as well as the sample sizes, the intra-class correlation, and the complexity of the models. We also compared results from several measures of model fit that have been suggested in the multilevel modeling literature (Nakagawa & Schielzeth, 2013; Raudenbush & Bryk, 2002; Snijders & Bosker, 1994). Preliminary results suggest that dominance analysis can be adequately used to determine the relative importance of predictors in multilevel models under certain conditions, and recommendations are made regarding the use of the procedure for this purpose.

MLM 3

**Poster 77: Effects of Multicollinearity on Parameter Estimates in Multilevel Growth Modeling**
Sang-Jin Kang, Yonsei University; Hyun-ah Ku, Yonsei University; Guemin Lee, Yonsei University

The issues of multicollinearity are well known to the researchers especially in regression modeling. In multilevel modeling, however, there exist only a few studies on the effects of multicollinearity (Kreft & deLeeuw, 1998; Sheih & Fouladi, 2003). Moreover, these are limited in two ways: First, all the studies investigated the effects of multicollinearity in cross-sectional analysis. Second, they all focused on the multicollinearity between the level-1 predictors even though multilevel model could employ the predictors from level-2 units. This simulation study investigated the properties of both fixed and random estimators given the multicollinearity among level-2 variables in multilevel growth modeling. More specifically, we examined the relative bias and precision of the parameter estimates of a two-level linear growth model in which both intercept and slope coefficient at level-1 model were specified as having different set of multiple predictors at level-2. Data generations were processed under the conditions designed by the combination of three conditioning variables. They were: 1) The size of correlations between level-2 variables modeling intercept at level-1 model, which included low (0.30), threatening (0.70), and definite (0.90). 2) The sample size at level-2 that included small (n=30), middle (n=50), and large (n=150). 3) The number of repeated observations within individuals that included small (T=3), middle (T=5), and large (T=9). The key results were that the multicollinearity greatly influences the precision but has little effect on the relative bias of the estimates of fixed and random parameters. The multicollinearity in the level-2 model for the intercept did not influence the estimates of regression coefficients modeling the slope.

MLM 4

**Poster 79: Mean Difference of Standardized Mean Difference in Fixed-Effect Meta-Analysis**
Rongwei Sun, University of Macao; Shu Fai Cheung, University of Macao

This study compared the influence of non-normality in raw data on the meta-analysis of Mean Difference (MD) and Standardized Mean Difference (SMD) for two independent group designs under the fixed effect model. Bias, Mean Square Error (MSE), Confidence Interval (CI), and the distribution and rejection rate of the Q statistic under different types of distribution of raw data were compa+AF75red. The results showed that in most conditions examined, the meta-analysis of MD performed better than that of SMD in the aforementioned aspects, even if the populations were moderately nonnormal. However, both MD and SMD performed unsatisfactory if the sample sizes were small and either population was exponentially distributed. The results suggest that adverse impacts due to non-normality in primary studies may not disappear even when the number of large studies is very large.

RES 1

**Poster 80: Three New Effect Size Measures in Structural Equation Modeling**
Brenna Gomer, University of Notre Dame; Ge Jiang, University of Notre Dame; Ke-Hai Yuan, University of Notre Dame

Effect size measures are crucial for quantifying differences and are a key concept behind type I and type II errors, but they are seldom studied in Structural Equation Modeling (SEM). While fit indices such as RMSEA may address the severity of model misspecification, they are not a direct generalization of commonly used effect size measures such as Cohen's d. Moreover, with a violation of normality and when a test statistic does not follow a non-central chi-square distribution, the square root of the non-centrality parameter, the RMSEA, and other measures of misfit that are defined through the test statistic are no longer valid. In this study, we propose three new effect size measures for SEM by generalizing the formulation of Cohen's d to SEM. There are several versions of such generalizations and we investigate which one is least affected by sample size and underlying

distribution and which is most sensitive to model misspecification. We examine their performance under violated distributional assumptions, varying sizes of model misspecification, and assorted sample sizes. We implement the commonly used normal maximum likelihood estimation and the more robust M-estimation to study these measures. Monte Carlo results indicate that a new measure following the robust method is much less affected by sample size and is more sensitive to model misspecification, and thus is recommended for researchers to report in publications.

SEM 1

**Poster 81: Data Deviations from Normality: Effects on Growth Curve Model Estimates**
Catarina Marques, Instituto Universitário de Lisboa; Maria de Fátima Salgueiro, Instituto Universitário de Lisboa; Paula C.R. Vicente, Universidade Lusfona &Instituto Universitário de Lisboa

In recent years there has been a substantial increase in longitudinal data collection, as well as in developing statistical methods and software to analyze such data. Latent Growth Curve Models (LGCM) are one of the very popular longitudinal statistical techniques: They allow individuals under analysis to have distinct growth trajectories over time. These patterns of change are summarized in relatively few parameters: in a LGCM with unconditional growth the main parameters of interest are 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 parameter structure imposes normality assumptions, the data analyst often faces data deviations from normality, implying small, moderate or even severe values for skewness and or kurtosis. A Monte Carlo simulation study was conducted in R and results are presented in order to investigate the effect of observed data deviations from normality on the estimates of the parameters of a latent curve model with unconditional linear growth. Different numbers of time points (3 to 6) and sample sizes (from 50 to 500) are considered. The effects of various levels of data skewness and kurtosis on the bias, on the mean square error and on the coverage of the model parameter estimates are assessed (Muthén & Asparouhov, 2012).

SEM 2

**Poster 82: Examining Model Fit in Confirmatory Factor Analysis with Ordinal Data**
Christine DiStefano, University of South Carolina; Grant B. Morgan, Baylor University

Fit indices are routinely used with covariance structural modeling evaluations. Information from fit indices can be used to inform research about the fit between a tested model and data, leading one to confirm (or refute) a hypothesized structure. While many recommendations and rules of thumb have been presented for evaluating fit indices, these guidelines are largely built from investigations with continuous, multivariate normal data and use of normal theory estimators (e.g., ML, GLS). However, estimators with robust corrections are prevalent in the literature. Additionally, with empirical studies, data may not be continuous (e.g., collected with a Likert scale) and also non-normally distributed. The purpose of this study is to examine the performance of fit indices under situations involving categorical data and non-normality. The research will vary factors of model misspecification, and sample size with a confirmatory factor analysis situation to determine which ad-hoc fit indices perform optimally under the tested conditions.

SEM 3

**Poster 83: Factor Score Regression for Components of the Social Relations Model**
Justine Loncke, Ghent University; Tom Loeys, Ghent University; Veroni Eichelsheim, Netherlands Institute for the Study of Crime and Law Enforcement

The Social Relations Model (SRM; Kenny & La Voie, 1984) with roles allows us to study how interpersonal dynamics manifest in families. When dyadic measurements are obtained from a

round-robin design, the SRM decomposes these measurements into a latent family effect, a latent actor and partner effect (at the individual level) and a latent relation-specific effect (at the dyadic level). Most SRM applications are limited to the exploration of the proportion of variance in the dyadic measurements that is explained by each of those SRM-effects, but they rarely identify the antecedents or consequences of those effects. Since simultaneous modeling relying on Structural Equation Modeling (SEM) may become computationally prohibitive in the small samples that are often seen in this setting, we consider several Factor Score Regression (FSR) methods: regression FSR, Bartlett FSR, regression FIML FSR and Bartlett FIML FSR. A simulation study is presented that compares the performance of the four methods with SEM in a complete-case setting as well as in the presence of missing data. In the complete case setting and in the presence of missingness, respectively, using the regression factor scores as explanatory variable and the regression FIML factor scores as explanatory variable, respectively, produce unbiased estimators with precision comparable to the SEM-estimators. When SRM-effects are used as dependent variables, Bartlett (FIML) FSR yields typically unbiased estimators (except for the family effect) that are less precise than the SEM-estimators. An empirical case study is presented that illustrates these results.

SEM 4

**Poster 84: Item Dimensionality and Number of Parcels on Parceling in SEM**
Li-Chung Lin, National Taiwan University; Yo-Lin Chen, National Taiwan University; Li-Jen Weng, National Taiwan University

Individual items can be aggregated to form parcels to represent indicators of latent variables in Structural Equation Modeling (SEM). SEM using parcels could yield less biased estimates and higher power in detecting misspecification than using items (Rhemtulla, 2016). In response to the scarcity of research on the effect of parceling with multidimensional constructs, Cole et al. (2016) conducted a simulation comparing the performance of two parceling strategies, facet-representative parceling and domain-representative parceling. The multidimensional construct was operationally defined as a second-order factor with three first-order factors representing its distinct facets. Facet-representative parcels combined items from the same first-order factor and items from different first-order factors were mixed to form domain-representative parcels. Beyond Cole et al., this study considered the number of parcels (Rogers & Schmitt, 2004) and multidimensional items commonly encountered in practice (Marsh et al., 2013). The number of parcels per second-order factor, number of items per first-order factor, factor loadings, item dimensionality, and sample size were manipulated to fully examine the effects of parceling strategies on SEM results. For conditions similar to Cole et al., where items reflected only one first-order factor and each second-order factor was indicated by three parcels, the findings from Cole et al. that facet-representative parceling resulted in more improper solutions, less biased structural parameter estimates, and larger standard errors than domain-representative strategy were replicated. Yet, differences between the two strategies diminished under multidimensional items or larger number of parcels for each second-order factor, yielding similar structural parameter estimates and standard errors.

SEM 5

**Poster 85: Model Evaluation with Small N and/or Large p.**
Lin Xing, University of Notre Dame; Ke-hai Yuan, University of Notre Dame

RMSEA and CFI are the two most widely used fit indices for evaluating the adequacy of a SEM model. They are typically computed via the likelihood ratio statistic TML. Under the normality assumption, TML approximately follows a chi-square distribution when the number of observations (N) is large and the number of variables (p) is relatively small. In practice, however, p can be rather large and N is always limited due to not having enough participants. Previous studies showed that RMSEA and CFI do not behave well with either a large p or a small N. RMSEA with established cutoff values tends to over-reject a correct model for small N, and so is CFI as the number of variables

increases. The poor behavior of RMSEA and CFI might be related to the problem in TML which tends to reject a correct model when N is small and/or p is large. With three recently corrected test statistics TMLe, this study focuses on the performances of RMSEA and CFI defined via the new statistics. Results indicate that the new RMSEA and CFI do perform better. They are much closer to their population values and the type I error rates are also close to the nominal level even for conditions with N as small as 2p.

SEM 6

**Poster 86: Three-Way Generalized Structured Component Analysis**
Seungmi Yang, McGill University; Ji Yeh Choi, McGill University; Arthur Tenenhaus, CentraleSupélec; Heungsun Hwang, McGill University

Generalized Structured Component A+AF78nalysis (GSCA) (Hwang & Takane, 2004, 2014) is a component-based approach to structural equation modeling, in which weighted composites or components of observed variables are used as proxies for conceptual or latent variables. GSCA has thus far focused on analyzing two-way (e.g., subjects by variables) data. We propose to extend GSCA to deal with three-way data that contain three different types of entities (e.g., subjects, variables, and occasions) simultaneously. The proposed method, called three-way GSCA, permits each latent variable to be loaded on two types of entities, i.e., entities in both second and third modes, in the measurement model. This enables to investigate how these entities are associated with a latent variable. The method aims to minimize a single least squares criterion to estimate parameters. An alternating least squares algorithm is developed to minimize this criterion. To demonstrate its empirical usefulness, three-way GSCA is applied to a part of the National Longitudinal Survey of Youth 1979-Children (NLSY79-C) data, a longitudinal study following up the biological children of female participants in NLSY79 every two years starting in 1986. In this example, we examine the relationships among three sets of three-way data, each of which is an array of subjects by variables by time points.

SEM 7

**Poster 87: The Single Indicator Method in Multilevel Structural Equation Modeling**
Soyoung Kim, Korea University; Youna Jang, Korea University; Sehee Hong, Korea University

The purpose of this study is to identify under which conditions single-indicators can be applied in Multilevel Structural Equation Modeling (MSEM). A simulation study was conducted with 2-level MESM having an independent variable and a dependent variable in the lower level. Both of these latent variables had six observed variables. The simulation conditions varied according to Intra-class Correlation (ICC), Number of Cluster, and Cluster Size. Analysis methods were the six indicators model, the single-indicator model with omega, and the single-indicator model with alpha. The bias and the convergence rate for each condition were evaluated. Results showed that, in the aspect of convergence, single-indicator models were superior to the all indicators model. For between-path coefficient, there were no differences between the single-indicator model and all indicators model when the cluster size was smaller. The bias decreased when the single-indicator method was used and worsened when ICC was smaller. For within-path coefficient, the bias of single-indicator method did not decrease greatly even when the cluster size or the number of cluster increased, but increased when ICC was bigger. In the single-indicator model, the method with omega and the method with alpha were almost the same. The discovery of the current research is as follows. In MSEM, in the case of between-path, the single-indicator model is better than the six indicator model when the cluster size is small. In the case of within-path, the standard error bias grows as the cluster size increases, thus using single-indicator model is not recommended when the cluster size is large.

SEM 8

**Poster 88: Gene-Gene Interaction Analysis Using a Generalized Structured Component Analysis Model**

Taesung Park, Seoul National University; Sungkyoung Choi, Seoul National University; Sungyoung Lee, Seoul National University; Heungsun Hwang, McGill University

Gene-Gene Interactions (GGIs) are known to be one possible explanation for lack of heritability in genetic association studies. While a number of approaches have been developed to identify GGIs, none of the approaches are able to account for biological structures in the analysis. For example, each gene contains many single nucleotide polymorphisms (SNPs). Most traditional approaches for detecting GGIs have only focused on SNP level interaction analysis. Unfortunately, SNP level interaction does not fully represent the GGI between genes. Here, we propose a novel statistical approach for GGI analysis, Hierarchical COMponent based Gene-Gene Interaction (HiCOM-GGI) analysis. The HiCOM-GGI method is based on Generalized Structured Component Analysis (GCSA). Unlike other previous methods for detecting GGIs, our HiCOM-GGI fits one large augmented model and evaluates the effects of both gene-level and SNP-level interactions. That is, for the given number of SNP sets, our COM-GGI model considers all possible two-way interactions into one model. Thus, higher order interactions can easily be incorporated into the model. In addition, HiCOM-GGI model provides much more accurate predictive results than other prediction models based on SNP level interactions. We illustrate HiCOM-GGI using the Age-Related Eye Disease Study (AREDS) data focusing on the body mass index (BMI) phenotype.

SEM 9

**Poster 89: Testing Structural Equation Models with Generalized R Squares**

Xiaofeng Steven Liu, University of South Carolina

Testing structural equation models traditionally uses model fit indices to examine how closely the variance covariance matrix implied by the model matches the variance covariance matrix based on the data (Bentler, 1986; Browne, 1984; Joreskog, 1977). A chi square statistic can be computed to test the differences in variance and covariance and compare the implied models with reference to the observed data. This paper offers an alternative strategy to test and compare structural equation models, similar to that in multiple regression analysis. The latter -- a special case of structural equation modeling -- uses R squares in model comparisons. The R square in regression analysis can be extended to multivariate outcomes in structural equation modeling. The variance of one outcome in a regression model becomes the generalized variance of multivariate outcomes in a structural equation model, which can be measured by the determinants of the variance and covariance matrices. A generalized R square can be derived from the determinants of the variance and covariance matrices for a structural equation model. In testing two SEM models, we can compare their generalized R squares. The SEM model with a significantly higher generalized R square is preferred over the competing SEM model with a lower generalized R square. A chi square can be used to compare the generalized R squares between two SEM models, and the chi square test is related to the tests based on the model fit indices.

SEM 10

**Poster 90: Facet- versus Domain-Representative Parcels with Multidimensional Constructs in SEM**

Yo-Lin Chen, National Taiwan University; Li-Chung Lin, National Taiwan University; Li-Jen Weng, National Taiwan University

Item parcels, represented as summation or average over items, have received increasing popularity in Structural Equation Modeling (SEM) when the goal of research aims at depicting the relations among constructs. When the constructs of interest are multidimensional encompassing multiple facets, facet- and domain-representative parceling can be used to construct parcels. Facet-

representative parcels contain items from the same facet, while items from different facets are combined to form domain-representative parcels. Cole et al. (2016) demonstrated graphically that the nature of multidimensional constructs from parcels differed depending on the strategy used and conducted a simulation to compare the results from the two strategies on a simple model with one multidimensional construct influencing one observed variable. This study extended the research of Cole et al. in three aspects. We algebraically derived sources of variation embedded in the constructs represented by the two types of parcels, adopted a more complex SEM model from Coffman and+AF102 MacCallum (2005) for generating artificial data, and considered items reflecting more than one facet from a multidimensional construct as frequently observed in real data. We replicated the results from Cole et al. with pure measures that the nature of constructs remained intact from facet-representative parcels but was contaminated for domain-representative parcels, and the parameter estimates from facet-representative parcels showed less bias. Yet, when items were affected by more than one facet, the constructs from both strategies were contaminated by the uniqueness of the subordinate facets, and similar parameter estimates and model fit were obtained from the two strategies.

SEM 11

**Poster 91: Efficient Scoring Method for Handwriting Constructed-Responses on Large-Scale Testing**

Tomoya Okubo, The National Center for University Entrance Examinations

In this study, an efficient scoring method for handwriting constructed-responses is proposed. It is a set of procedures including cluster analysis on the basis of electronic text data generated by intelligent character recognition. It aims to reduce cognitive loads of human raters by sorting responses according to their similarity. It is expected that the scoring sorted responses are easier than random responses. This idea is a big change in scoring strategy, since we aim at supporting human raters, while previous studies pursue accuracy of their scoring model. An advantage of our proposing method is that it does not require 100% accuracy of the intelligent character recognition. Even incomplete electronic text data can be used. In this study, we validated effects of our proposing method using a large-scale test. The result shows that the proposed scoring method is useful even when digitized text data obtained by the intelligent character recognition is not perfect. It also showed that scoring similar responses successively reduces loads of judgments in human raters and increases speed of scorings.

SML 1

**Poster 92: Accuracy of Linking VR-12 and PROMIS Global Health Scores in Clinical Practice**

Brittany Lapin, Cleveland Clinic; Tyler Kinzy, Cleveland Clinic; Nic Thompson, Cleveland Clinic; Irene Katzan, Cleveland Clinic

The study aim was to examine the accuracy of general health cross-walk tables in a clinical population of spine patients. Recently published tables (Schalet et al, 2015) link scores from a commonly used measure, the Veterans Rand 12-Item Health Survey (VR-12) to the 10-item Patient Reported Outcome Measurement Information System (PROMIS) General Health scale metric for both Physical (PCS) and Mental Component Scores (MCS). We assessed accuracy of administered PROMIS and VR-12 scores with scores predicted by cross-walks in 4,606 adult patients seen in spine clinics at Cleveland Clinic during the past year. Accuracy of linking was evaluated using Pearson correlation, Intraclass Correlation Coefficients (ICC), and mean and standard deviation of score differences. Bland-Altman plots graphically assessed levels of agreement. Consistency in discrimination between pain severity, depression, and characteristics was assessed. Bootstrap methods estimated linking bias across varying sample sizes. Actual and cross-walked PROMIS scores showed moderate correlation (ICC: 0.73 (MCS), 0.81 (PCS)), with Bland-Altman plots suggesting smaller differences between scores in patients with lower and higher general health.

Significant discrimination between patient subgroups was demonstrated reliably by both actual and estimated scores. Bootstrapped resamples indicated minimal bias at n=200 (95% CI for mean difference, MCS: -1.38-0.60; PCS: 0.39-1.93). Reverse linking (PROMIS to VR-12) through linear interpolation exhibited similar results. VR-12 and PROMIS Global Health scores can be accurately linked within a spine patient population. Linked scores for 200+ patients can be used in comparative effectiveness research and for comparing results across studies, increasing the generalizability of these measures in outcomes research.

VAL 1

**Poster 93: A Meta-Analysis of Stability of Aggressive Behavior**
Chiungjung Huang, National Changhua University of Education

The purposes of the present meta-analysis were twofold: (a) to estimate the overall longitudinal stability of aggressive behavior; and (b) to examine the moderator effects on the stability of aggressive behavior. Longitudinal studies (at least a 1-year interval) assessing aggressive behavior during school years were included. As the distribution of correlation coefficients was skewed, the stability coefficient r was converted to a normalized stability coefficient using Fisher's r to Zr transformation equation. The mean and confidence intervals of Zrs were computed and then transformed back into weighted mean stability coefficients. Inverse-variance weighting was used to compute the average effect sizes. Forty-four independent samples involving 11,261 participants were identified. Coding multiple effect sizes for studies with more than two waves of data collection yielded 63 test-retest correlations. The weighted mean stability coefficient was .48. The stability coefficients for aggressive behavior are heterogeneous. Moderators, time interval, age at initial testing, gender composition, ethnicity composition, sample risk status, and the cohort were entered into the weighted regression model to explain the variability of the stability of aggressive behavior. Time interval between waves was the only significant moderator of aggressive behavior stability. The stability coefficient decreased as the time interval between rounds of data collection increased when holding all other moderators constant. However, the six moderators together did not explain a significant amount of effect size variance.

VAL 2

**Poster 94: Item-Score Reliability in Real-Data Sets**
Eva A. O. Zijlmans, Tilburg University; Jesper Tijmstra, Tilburg University; Andries L. van der Ark, University of Amsterdam; Klaas Sijtsma, Tilburg University

Reliability is usually estimated for a total score, but it can also be estimated for item scores. Item-score reliability can be useful to assess the repeatability of a set of item scores in a group. Three methods to estimate item-score reliability are discussed, known as method MS, method λ6, and method CA. The item-reliability methods are compared to four well-known and widely accepted item indices, which are the item-rest correlation, the item factor-loading, the item scalability, and the item discrimination. Item-score reliability in real-data sets is estimated by the three methods to obtain an impression of the values to be expected in other real-data sets, and the relation between the three item-score reliability methods and the four well-known item indices (e.g., item-rest correlation, item factor-loading, item scalability, and item discrimination) are investigated to see whether they identify the same items as weak or strong with respect to their placement in a scale. A recommendation for a lower bound for item-score reliability to be used in item analysis is recommended.

VAL 3

**Poster 95: The Unidimensinality of Social Desirability in the Polish Population**
Ewa Witkowska, Maria Grzegorzewska University; Krzysztof Fronczyk, University of Warsaw

In our previous studies on social desirability structure in the Polish population (Fronczyk & Witkowska, submitted 2017) the unidimensional structure of the social desirability construct in the Polish population was revealed. This verification of the social desirability structure was performed in three stages, using the originally generated poll of items. In each stage, the number of items was gradually reduced using factor analysis. Initially, in the first stage of the study, two dimensions were detected, but in the two subsequent stages the second dimension represents acquiescence, not a real psychological meaning. Finally, a one-dimensional questionnaire was obtained. Such a structure is contradictory to the dimensions proposed by Paulhus (1984, 2002) or Ramanaiah et al. (1977), supporting the existence of two social desirability dimensions. However, the result from Ramanaiah et al. (1977) later proved indicate that the two-dimensional structure was strongly influenced by the responding style. Our failure to replicate the two-factor structure of social desirability was then congruent with Ramanaiah and Martin's findings (1980). This paper is aimed in presenting further evidence from the Rasch model and parallel analysis supporting unidimensional nature of social desirability among Poles.

VAL 4

**Poster 96: Evaluating the Utility of 101-Point Numerical Rating Scales**
Hsin-Yun Lee, National Taiwan University; Li-Jen Weng, National Taiwan University

Would a scale with a large number of response categories elicit scores more reflective of participant predisposition and make the scale closer to a continuous measure? The current study investigated the use and effectiveness of a numerical rating scale composed of 101 points (NRS-101). Participants were asked to respond to each item with an integer between 0 and 100. A total of 1,929 college students completed two subscales of the Teacher Attitude Test both on the NRS-101 and on Likert-type scales that differed in scale formats. The number of response categories ranged from 3 to 9. Anchor labels on scales were provided for each response option, for the endpoints and the middle category, or for the endpoints only. Examination of the actual numbers used in the NRS-101 indicated that almost all the participants offered responses in multiples of 10 or 5, treating it as a discrete scale. Scores from NRS-101 were highly correlated with responses from the Likert-type scales, especially for scales with heterogeneous participant responses. The NRS-101 tended to yield higher internal consistency reliability estimates than Likert-type scales yet the differences were trivial for scales exhibiting larger variation. These findings suggested that although the NRS-101 seemed closer to a continuous scale, respondents might not perceive it as continuous. The superiority of NRS-101 over Likert-type scales depended on the degree of heterogeneity of participant responses on the scales. Psychometric properties of scales with homogeneous participant responses might be improved by using NRS-101.

VAL 5

**Poster 97: Estimating CSEM for Test Scores with Balanced and Unbalanced Data**
Hyejin Kang, Yonsei University; Nana Kim, Yonsei University; Guemin Lee, Yonsei University

Testlets are defined as small subsets of a larger test (Wainer & Kiely, 1987). Tests composed of testlets can be classified into "balanced" and "unbalanced" data structure. The balanced data structure indicates that each testlet has the same number of items while the unbalanced data structure does not satisfy this condition. Under Generalizability Theory (GT), three approaches can be considered for estimating the reliability of test scores with testlets. One is the item-based approach (p×I design) which uses individual item scores and ignores the dependency among items within testlets. Another approach is to aggregate item scores within testlets and treat those as polytomous items (p×H design). The last is to reflect the structure of items within testlets in the

model (p×(I:H) design). Park, Lee, and Lee (2011) found that reliability estimates were influenced by the imbalance in the data structure. They reported that as the degree of imbalance increased, the degree of underestimation/overestimation of reliabilities increased. It might be reasonable to expect that the degree of imbalance would also affect estimates of Conditional Standard Error of Measurement (CSEM). This is the main motivation for conducting this study. This study aims to investigate the effects of balanced and unbalanced data structures on the CSEM estimates for test scores with testlets using simulation techniques. We compare the results of three GT approaches in the CSEM estimates of test scores with testlets, especially focusing on the effects of the degree of imbalance across testlets and levels of dependence within testlets.

VAL 6

**Poster 98: The Effect of Different Item Weightings on Reliability Coefficients**

Hyesung Shin, Yonsei University; Guemin Lee, Yonsei University

A composite score is a single index obtained by summing various separate measures of a person (Dunnette & Hoggatt, 1957) and differential weighting is an important topic when forming a composite score (Marilyn & Julian, 1970). However, differentially weighted composite scores have been criticized because they were found to be ineffective. Odell (1931) and Lara et al. (2006) concluded that weighting was not worthwhile as regards improving the reliability of a test. To be specific, Ruch & Meyer (1931) found that weighting by item difficulty lowered the reliability. Petthoff & Barnett (1932), Kim & Ro (1999), Yang (2007) and Park & Kang (2011) also found that there was no significant difference in composite score and reliability when using expert-based weights compared to equal weights. However, different weighting is still widely used in test practices, especially for high-stake tests such as college admission tests. The purpose of this study is to compare the reliability of composite scores using different weights under a generalizability theory framework. Four different weighting methods are considered. The first method is equal weights, which corresponds to no weights (Marilyn & Julian, 1970). The second method is weighting on the basis of difficulty, which is most popular (Marilyn & Julian, 1970; Kim & Ro, 1999). The third is weighting by item discrimination, since discriminating students is the goal of weighting. The last one is weighting at random. A simulation study will be conducted. Response data will be generated with three different test lengths: 20, 30 and 40 and different weighting methods will be applied for estimating the reliability.

VAL 7

**Poster 99: Investigation of the SEM for Standard Setting Using Multivariate Generalizability Theory**

Jiyoung Jung, Yonsei University; Heewon Yang, Yonsei University

The adequacy of cut-off scores is very important for reporting achievement levels and analyzing trends for criterion-referenced assessment. Various standard setting methods have been devised for setting the cut-off score. The most frequently used methods for standard setting are the Angoff and Bookmark methods. The cut-off score can be determined in different ways depending on the standard setting method, so examinees that obtained the same score can be grouped into different categories. Thus, comparing and assessing the reliability of the cut-off score can be one indicator when choosing standard setting method applied to test results. The research objectives for this study are as follows. First, to investigate sources of error and generalizability coefficients for comparison to the Angoff and Bookmark methods under multivariate generalizability theory framework. Second, to compare the standard error of measurement (SEM) of cut-off scores between the Angoff and the Bookmark methods. Third, to analyze consistency within the Angoff and the Bookmark methods using classification consistency and accuracy. In this study, cut-off scores were obtained from standard setting for the 5th grade Korean Educational Longitudinal Study (KELS) conducted in 2013, and the MGENOVA computer program were used for analysis. The

results of this research can be summarized as follows. First, the error source and effect size of each error source differ depending on the method of standard setting across on univariate and multivariate generalizability theory. Second, the Angoff method showed higher standard error of measurement of cut-off scores than the Bookmark method. Third, consistency within method was higher with the Bookmark method.

VAL 8

**Poster 100: Reverse-Worded Items and Their Threat to Validity: Evidence and Remedies**
Karolina Świst, Educational Research Institute; Marek Muszyński, Educational Research Institute & Jagiellonian University

The data from the Polish Follow-up study on the Programme for International Assessment of Adult Competencies (postPIAAC) include data on personality measures, e.g. BFI-S (Gerlitz & Schupp, 2005) short questionnaire based on the Big Five personality framework. Preliminary analysis (Palczyńska & Świst, 2016) shows that theoretical BFI-S structure (five-factor orthogonal) does not hold in the Polish sample and that the psychometric quality of the scale is only adequate. Lack of structural validity was especially pronounced in mixed-worded scales as it is often found in the literature (Swain, Weathers & Niedrich, 2008; Weijters & Baumgartner, 2012; Weijters, Baumgartner & Schillewaert, 2013). We attributed those problems to the presence of extreme and midpoint response style in the data, presence of which leading to the inconsistent responding on reverse-worded items, which introduces higher systematic and random measurement error (DiStefano & Motl, 2006; Marsh, 1996). He and van de Vijver (2016) approach is used to detect response styles and their socio-demographic correlates, (van Vaerenbergh & Thomas, 2013; Baumgartner & Steenkamp, 2001; Bartram, Inceoglu & van de Vijver, 2014; Rammstedt, Goldberg & Borg, 2010). Age correlates positively with response style, while educational attainment and numeracy correlates negatively, which confirms Meisenberg and Williams (2008) results. Controlling for the response styles improves the structural validity of BFI-S. Closing remarks concern use of reversed-worded items in large scale surveys, especially with participants of mixed educational attainment.

VAL 9

**Poster 101: Psychometric Features of the Final BDS Assessment**
M Khalid, Cardiff University; Sheila Oliver, Cardiff University; Ilona Johnson, Cardiff University

The validity and reliability of test scores are important concerns in education setting. There are numerous measures to demonstrate these crucial aspects of any high stake test. In the present study, a number of explanatory psychometric analysis were used to assess the reliability and validity of BDS final exam administered by the School of Dentistry, University of Cardiff. We investigated the inter rater reliability to examine how much homogeneity was in the ratings given by judges. This included the removal of outliers to see the impact on the cut score, reliability of assessment scores, decision consistency and accuracy, standard error of measurement of assessment scores and cut scores, difficulty of the questions, differential performance of students and judges across topics and comparability of standard setting approaches. The examination involved three papers, each with 15 multiple short answer questions with a total of 10 points per question. The results showed the reliability of assessment score was 0.876 with a SEM of 2.06 and inter rater reliability across judges was 0.92 with a SEM of 0.14. The results of the study though suggested a trivial differential performance of judges across the topics but there was no significant impact on the overall cut score. This study illustrated classification of examinees was consistent and accurate having reliability of 0.96. The results of the study supported that we established credible, defensible, and acceptable passing or cut off scores for our assessments. The analyses enabled us to provide constructive feedback to the judges about their ratings and consistency.

VAL 10

**Poster 102: Mapping the Cognitive Process of Rubric Rating**
Nicholas Curtis, James Madison University; Allison Ames, James Madison University; Madison Holzman, James Madison University

Rubric-based performance assessment is promoted as more realistic than other assessment methods (Wiggins, 1991). Limitations of rubrics are well-documented (e.g., cost, time, effort), and because rubrics require human raters, they are prone to rater error. Despite this, proponents contend that interpretations are strong since rubrics often allow more complex and direct assessment of real-world performance. Attempting to demonstrate that subjectivity is not an issue, statistics such as inter-rater reliability are used. Unfortunately, reliability is necessary, but not sufficient, validity evidence for such a claim. Raters giving the same score do not necessarily use the same logic and decision-making to assign scores. To the extent to which raters do not use the same decision-making process to assign scores, validity is compromised. Methods to address decision making processes include rater training and think-aloud protocols. These methods have provided indirect evidence of thought processes in responding to rubrics; however, they fall short of providing strong, direct evidence. To address the lack of cognitive process evidence, we developed a new method of scoring performance assessments. This method explicitly guides raters to think and respond in ways more similar to expert raters. Instead of a traditional rubric, raters receive a series of multiple-choice questions to match expert-guided cognitive rating processes. Computer adaptive logic is employed such that subsequent questions are presented based on answers to previous questions. Consequently, raters are required to use a focused progression of logic. Results suggest the new method provides stronger validity evidence than obtained using traditional rubrics.

VAL 11

**Poster 103: Estimating Domain Scores Under IRT and Generalizability Theory Approaches**
Youkyoung Oh, Yonsei University; Seonghyun An, Yonsei University; Guemin Lee, Yonsei University

Domain scores can be defined as scores on a set of items for certain content that provided more comprehensive information than scores on individual items (Bock et al., 1997). These scores were computed by total possible points examinees earned in a domain. If a large number of items in a domain are administered, the estimation of a domain score would be precise and reliable. However, in practice, the number of items used in one content area is not enough to reach the desired reliability level (Shin, 2007). Thus, four estimation methods would be considered for minimize the problem (Bock et al., 1997). This study aims to investigate the relative appropriateness for IRT and G-theory in estimating domain scores. Limited researche has been conducted on estimation of reliable domain scores (Bock et al., 1997; Harris et al., 2000; 2002; Harris et al., 2003; Shin & Lee, 2010). Domain scores will be estimated with different number of items in domains (13, 26, 51) and five methods will be employed (observed scores, IRT-EAP, GT-Searle, GT-Cronbach, GT-Jarjoura). The specific research objectives of this study are as follows: 1) Which method performed best among the five methods, 2) What effect does the number of items taken from a domain have on estimating the domain score. In primary research, regardless of the number of items, the IRT-EAP method outperformed the other four methods. There were no significant decreases in the absolute difference and MSE from 13 to 26 items.

VAL 12

**Poster 104: Formal Modeling of Projective Techniques and Their Validation**
Yury Chernov, Institute for Handwriting Research

Projective techniques in general and the handwriting psychology in particular have certain unique qualities for the psychological assessment. The major are a wide coverage of personal characteristics and the exclusion of social desirability. Their weakness is the insufficient validation

that makes them controversial. The current presentation describes a formal model of the handwriting analysis and a new approach to the validation (applicable to other projective methods as well). Most of the known studies on this topic suffer from serious methodological defects: they are based on manual informal handwriting evaluation, typically involve insufficient experiment data, analyse very limited number of the handwriting characteristics, and use inappropriate statistical methods. AF106The presented model (HSDetect) includes statistically evaluated relations between 786 handwriting signs and 370 personality traits and behaviour patterns. The objectivity and reliability of the approach was assured in numerous studies by means of expert procedures, statistical (e.g. Cronbach-Alpha) checks and experiments with a graphics tablet. The validation was examined against well-known personal tests like NEO-FFI, 16PF-R, EQ-i, PVQ and some others. The presented system allows adequate modelling and evaluation of test dimensions by a set of psychological traits, rather than just one trait, what was typical with a manual procedure. The results show statistically significant agreement between the psychological tests and the handwriting analysis on most test dimensions. That is promising and open new opportunities for the modernization and adaptation of the handwriting psychology to the quickly changing handwriting habits.

VAL 13