IMPS 2019

International Meeting of the Psychometric Society

Santiago de Chile

Abstract Book: **Posters**

Updated on June 1, 2019

Contents

1 Item set assembly and psychometric properties of set-based tests

Usama Ali, *Educational Testing Service, United States*
Manfred Steffen, *Educational Testing Service*

Reading comprehension sets have long formed the foundation for many verbal or language tests. In an attempt to making the assessment of reading comprehension more authentic, test developers have tended toward the use of longer reading passages. Accordingly, the number of associated items tends to increase as well. As a result the more authentic the reading comprehension task is purported to be, the larger the number of items associated with the passage. Additionally, in order to reduce any context effects, the set members are delivered in a fixed sequence. While such assessments are quite authentic, they introduce two significant concerns. First by virtue of being related the items associated with reading sets do not produce independent responses and thus grossly overestimate the reliability of such assessments. Second, for testing programs that use a nearly continuous delivery model, items/sets are frequently reused. Reuse of large numbers of related items that appear in a fixed sequence introduces an even greater security risk is that the key sequence can be easily catalogued and shared. In this study our goal is to investigate different ways to treat item sets as pools of items from which a subset are selected at delivery time. The study addresses the guidelines for member sampling within item sets to improve the test security and assess the functional size of a set and whether a single optimal subset exists. The studied methods are widely applicable to a wide range of testing programs that are considered set-based assessments.

2 Estimation of transfer effects for the standardized Austrian matriculation exam

Philipp Gewessler, *Austrian Federal Ministry of Education, Science and Research, Austria*
Jan Steinfeld, *Austrian Federal Ministry of Education, Science and Research*
Michael Themessl-Huber, *Austrian Federal Ministry of Education, Science and Research*

The Austrian matriculation examination in mathematics is standardized nationwide and is required for accessing institutes of higher education. After the exam the items are published online and can be used by subsequent cohorts as exercise material. Publication of items means that they cannot be reused. Therefore a large number of candidate items is piloted every year, creating an item pool from which items are drawn based on theoretical and empirical considerations. To guide item selection by domain experts, a variety of item characteristics are provided. In order to facilitate the item selection, expected solution probabilities for each item in the high-stakes test are predicted using the shift of latent student population distributions between pilot test and high-stakes exam for each cohort. Due to the large number of test takers, joint estimation is computationally infeasible. Instead, the estimation of the parameters is split into 3 steps. First, all parameters in the pilot studies are jointly estimated using a Bayesian 1PL IRT model, where the parameters of the student population distributions are freely estimated for each cohort. Second, a 1PL IRT model for each high-stakes exam is computed assuming invariance of the item difficulties. Third, the means of the student populations are modelled in a Bayesian meta-regression, resulting in an estimate for the transfer effect between pilot tests and high-stakes exams. The shift in population means is then predicted for the current student cohort and accuracy of the results is evaluated by comparing the predicted probabilities to the observed solution frequencies.

3 The antecedents of trajectory of maternal educational expectations for their children

Surina He, *Beijing Normal University, China*
Xiaolin Guo, *Beijing Normal University*
Tingdan Zhang, *Beijing Normal University*
Tiantian Bi, *Beijing Normal University*
Huan Qin, *Beijing Normal University*

From the developmental perspective, the educational expectations of children and their parents were variable over time. Although many researches documented the growth of children's educational expectations (Mello, 2008, 2009), few studies investigated the trajectories of parental educational expectations. For the formation of parental educational expectations, recent studies increasingly concerned with the role of children's educational expectations and academic outcomes (Briley, et al., 2014). Thus, in present study, we aimed to investigate the effect of children's educational expectations changes and their academic achievement on trajectory of maternal educational expectations. In this study, we used four waves data of 2704 Chinese children and their mothers. The results of latent growth curve models revealed that the trajectory of maternal educational expectations showed linear growth in four times. After controlling some demographic variables, the results of structural equation models showed that the initial level of children's educational expectation significantly influenced the initial level of maternal educational expectations, meaning that mothers whose children had higher educational expectations tended to

have higher educational expectations for their children. Moreover, the slope of children's educational expectation positively affect the slope of maternal educational expectations, which meant that the changing rate of maternal educational expectations were influenced by that of their children's educational expectations. Furthermore, children's academic achievements also had positive effect on initial level of mothers' educational expectations and had negative effect on its changes rate.

4 Mechanisms of scientific experiments, interests, self-efficacy on achievements using SEM

Chaochao Jia, *Beijing Normal University, China*
Tao Yang, *Beijing Normal University*

Scientific literacy plays an important role in social and economic development. In this research, two theoretical models of multiple mediation effects are established between scientific experiments (demonstrative experiment and hand-on experiment), scientific interests, scientific efficacies and scientific achievements, and the databases of the Chinese National Assessment of Education Quality (NAEQ) are used for analysis of both science test and questionnaire. The results show that physical hand-on experiment, biological demonstration and hand-on experiment can significantly positively predict achievements, while physical demonstration experiment significantly negatively predicts achievement. The mediation models are almost indistinguishable between physical and biological disciplines, while the models change greatly with the mediator variables (scientific interest and efficacy) reversed the action directions. Advice is given for science teachers that more attentions should be paid on students' hand-on experiments rather than demonstration one. The paths of the mediation models are examined, which revealing efficacy has more influence on achievement than interest, and the direction of the two mediators is from interest to efficacy, rather than from interest to efficacy. Students' interest of science, which can be influenced by experiments, can significantly positively predict science learning efficacy and then influence the achievement of science.

5 The Brief Resilience Scale: An IRT analysis of longitudinal data

Brooke Midkiff, *Duke University, United States*

This study uses longitudinal data from 4 institutions across 5 years to conduct an IRT analysis of the Brief Resilience Scale (Smith et al., 2008). This research fills gaps in the literature around the psychometric properties of the Brief Resilience Scale, utilizing IRT to re-examine the factor structure and reliability determined by Smith et al. (2008). This study also re-examines test-retest reliability using a larger sample over a longer time than previously used, convergent validity with both the Brief Coping Scale (Carver, 1997) and the Life Orientation Scale (Scheier, Carver, & Bridges, 1994), and discriminant validity using the Positive and Negative Affect Schedule (Watson, Clark, & Tellegen, 1988). Additionally, this study provides analysis of local dependence between items and DIF between known subgroups including those by gender and underrepresented minority student status. The sample consists of 2,916 participants who completed the Brief Resilience Scale (Smith et al., 2008) as part of a multi-year, multi-institutional research project. Initially 2,916 students completed the scale in the first wave, with attrition falling to 950 in the intervening years, and additional participants joining the study in the final wave, with an average of $<$1% of missing data within panels (smallest panel size, n=848). Responses were analyzed for careless responding based on Mahalanobis distance (Weber, 2010) and coefficient of variation (Roßmann, 2017), yielding N=6197 total observations across panels. This is the first research with sufficient, real-life data to fully analyze the psychometric properties of the Brief Resilience Scale using IRT methods.

6 Chilean Parental Burnout and Gender. Results from the international parental burnout investigation.

Pablo Pérez-Díaz, *University College London, United Kingdom*
Daniela Oyarce Cádiz , *Universidad Autónoma de Chile*

Currently, the most extensive effort for understanding parental burnout is being led by researchers in more than thirty-five countries. The present study analyses the Chilean dataset of the IIPB regarding parental burnout and gender differences. For such purpose, we compared fathers and mothers through the parental burnout measures (i.e., Emotional exhaustion, Contrast with the previous parental-self, Parental Saturation, Emotional distancing, and Parental Burnout total score). All mean comparisons resulted two be significantly higher for women when compared to men but Emotional distancing ($p < .05$). Furthermore, we tested a univariate linear regression model which included parental function, independent and interdependent self-image, and goal & values transmitted by parents to children. The proposed model explained 16.7% of the variance for the criterion variable (R^2-adjusted). We discuss the results at the light of possible public policies on parenthood as this was the first study assessing the construct of parental burnout in the country.

7 Modeling non-compliance in the randomized response technique using unrelated questions

Fabiola Reiber, *University of Tuebingen, Germany*

Randomized response techniques aim at increasing the validity of measuring sensitive attributes by ensuring permanent anonymity protection of respondents. Because single responses are not conclusive of the respondents' status, respondents are assumed to be more willing to honestly disclose sensitive information about themselves. Though this is indeed more likely than in conventional direct questioning, instruction non-compliance, like giving self-protecting responses, is still possible. In the present work, a new Randomized Response Model is developed, building on two existing models, namely the Unrelated Question Model (UQM, Greenberg, Abul-Ela, Simmons, & Horvitz, 1969) and the Cheater Detection Model (CDM, Clark & Desharnais, 1998). The design of the UQM, which is characterized by its psychological acceptability for study participants and its favorable statistical properties, is combined with the CDM's assumptions concerning instruction non-adherence. We present explicit formulas to calculate the parameters of the resulting model, provide remarks on power and optimal choice of the design parameters and show how the empirical adequacy of this extended model can be tested.

8 Gender-based DIF on the Child Behavior Checklist in autism

Hillary Schiltz, *Marquette University, United States*
Brooke Magnus, *Marquette University*

Few studies have examined the psychometric properties of questionnaires among individuals with autism, a lifelong neurodevelopmental disorder characterized by difficulties with social engagement that is often accompanied by comorbid symptomology. Moreover, given the lower prevalence and unique presentation of autism in females, there is even less evidence to support the utility and performance of these measures across gender. Data on multiple self- and proxy-report measures were gathered from the National Database for Autism Research (NDAR), an NIH-funded data repository. Included participants met criteria for autism on the Autism Diagnostic Observation Schedule. We examined the factor structure of one of the most commonly used parent-report measures of emotional and behavioral functioning, the Child Behavior Checklist (CBCL). After identifying the factor structure, we used a multiple indicator multiple cause (MIMIC) model to test each subscale item for gender-based differential item functioning (DIF). Items with low factor loadings, highly correlated residuals, and low endorsement were taken into consideration and revised, such that final model for each CBCL subscale exhibited good fit. Items on the Rule Breaking subscale showed very little variability; thus, this factor was excluded from analyses. Gender-based DIF was found on the Social Problems subscale for items reflecting peer interactions, on the Aggression Problems subscale for items related to emotion dysregulation, and on the Thought Problems subscale for items reflecting hoarding behavior. This study has implications for future research examining the etiology, trajectory, and ultimately treatment of symptoms for autistic males and females.

9 Psychometric investigation of executive functioning: Student task and teacher report

Jessica Sperling, *Duke University, United States*
Lorrie Schmid, *Duke University*
Victoria Lee, *Duke University*
Megan Gray, *Duke University*

Executive functioning (EF) is a necessary prerequisite to a broad set of positive well-being and academic outcomes (McCelland et al., 2007). However, precise measurement of these behaviors is an open research question (Best & Miller, 2010; Willoughby et al., 2016). The purpose of this aspect of the study is to assess the psychometric properties of EF measures, to examine if EF is unidimensional, and, to determine the relationship the student tasks and teacher reports relating to EF. This study includes child neurobehavioral tasks completed within their current (Kindergarten or 1st grade) school year (n = 123). The three distinct tasks assess attention, working memory (backward digit span), inhibitory control (go/no-go), and planning and judgement (tower of London; Mueller & Piper, 2014). Teachers reported on child EF behaviors and attitudes using the Dimensions of Mastery Questionnaire (DMQ; Morgan et al., 2019). We explored the internal consistency of multiple indicators from each of the behavioral tasks, reviewed the correlations between the behavioral tasks and teacher report, and conducted confirmatory factor analyses (CFA) to ascertain dimensionality. Originally, we anticipated constructing a single factor; however, Willoughby's et al. (2016) article indicated a single factor may not be appropriate. Therefore, we conducted a bifactor CFA to account for variability in the specific measures and general structures. Taken altogether, this allows us to examine the interplay of EF in children of this age. Future work will look to determine how EF is related to social emotional competence and to determine changes to EF over time.

10 Modeling household food insecurity with a polytomous Rasch model

Victoria Tanaka, *University of Georgia, United States*
George Engelhard, Jr., *The University of Georgia*
Matthew P. Rabbitt, *U.S. Department of Agriculture*

The Household Food Security Survey Module (HFSSM) is an 18-item scale created and maintained by the U.S. Department of Agriculture (USDA) that measures food insecurity in the United States. The HFSSM includes ten items that reference food hardships among adults in the household, and eight items that reference food hardships among children. The scale was created and maintained using a dichotomous Rasch model (Engelhard, Engelhard, & Rabbitt (2017). However, the item responses that are collected for nine of the items are polytomous that are later dichotomized for creating the final scale. In 2006, the Committee on National Statistics (CNSTAT) reviewed the HFSSM and the USDA's procedures for measuring food insecurity. They suggested modeling polytomous item responses with a polytomous model instead of dichotomizing item responses (Wunderlich & Norwood, 2006). The purpose of this study is to explore modeling polytomous HFSSM items with a partial credit model, building on Nord's (2012) work on the partial credit model and the HFSSM. The polytomous Rasch model will be compared to the dichotomous Rasch model currently used by the USDA. Challenges and implications of modeling polytomous item responses will also be considered. The HFSSM is an important tool for understanding the prevalence of food insecurity in the U.S. and influencing national policy decisions. The proposed study will provide greater insight into the psychometric properties of the HFSSM and the measurement problems encountered when measuring food insecurity.

11 Research on the learning burden of primary school students

Ping Zhang, *Beijing Normal University, China*

The conflict between the individualized learning needs of different types of students is one of the obstacles to the implementation of the burden reduction policy. The present study focused on learning engagement and learning burden inside and outside the school through cluster analysis. This study examined the learning profiles of 8973 students from 44 primary schools in Beijing. The cluster analyzes revealed four distinct learning burden profiles, and they were compared in association with learning attitude, learning method, interpersonal relation, parental expectations, and academic achievement etc. Cluster profiling enables teachers to have better understanding of their students' burden so that they can apply effective teaching strategies to reduction their learning burden. The purpose of this study is to explore the characteristics of different groups of primary school students, in order to guide the reduction of learning burden in practice. Thus providing effective burden scheme according to different category students, enhance the overall efficiency of the education.

12 Evaluation of three IRT-based classification consistency indices

Jinmin Chung, *University of Iowa, United States*
Seohee Park, *University of Iowa*

Many assessments are utilized for classifying examinees such as pass/fail or more than two performance levels. This classification should be precise because it can affect examinees' future career or learning procedures. To evaluate the precision of classifications, classification consistency (CC) is commonly used, which is the likelihood that examinees are classified into the same categories over replication (Lee, 2010). Currently, two methods of estimating CC based on IRT (a. Rudner approach and b. IRT-recursive approach) have been used and researched. These methods require the estimates of latent trait score and item parameters. The IRT parameter estimation is usually executed through Maximum Likelihood (ML) estimation. However, ML estimation for IRT faces convergence issues because of small sample size or short test length. To supplement this issue, psychometricians have been investigating Bayesian estimation. Align with popularity of Bayesian, Wainer, Wang, and Skorupski (2005) proposed a new estimation method of CC by using outputs from Bayesian estimation, known as Poster Probability of Passing (PPoP). Although the decent number of previous studies have explored comparison between the methods of estimating classification, any research did not include PPoP. Considering this, the current study will investigate how differently three estimation methods of classification performance in non-common situations such as non-normal distribution of latent trait, small same size, small test length, and large number of parameters in IRT model.

13 Markov chain Monte Carlo methods for inferring dimensionality of diagnostic models

Ying Liu, *University of Illinois at Urbana Champaign, United States*
Yinghan Chen, *University of Nevada, Reno*
Steven Culpepper, *University of Illinois at Urbana Champaign*
Yuguo Chen, *University of Illinois at Urbana Champaign*

Recent research considered exploratory diagnostic models (DMs) to infer the structure underlying multivariate binary response data. Exploratory models broaden the applicability of DMs to applied research, but a limitation of existing methods is that the dimensionality of the latent space is assumed known and fixed. Bayesian inference about model dimensionality is complicated by the fact that the number of parameters changes with the number of attributes. We consider two Markov Chain Monte Carlo (MCMC) methods for inferring the dimensionality of the exploratory deterministic inputs, noisy "and" gate (DINA) model: 1) reversible jump MCMC; and 2) the Dirichlet process. We implement a simulation study to evaluate the performance of the MCMC methods and discuss the relative merits of each procedure for inferring the number of attributes for the exploratory DINA model.

14 IRT and CDM analyses using international assessment data

Diego Luna-Bazaldua, *National Autonomous University of Mexico, Mexico*
Young-Sun Lee, *Teachers College, Columbia University*

There has been recent growth in the number of new diagnostic psychometric methods, which either expand the Classical Test Theory (CTT) or the Item Response Theory (IRT) frameworks or propose new latent variable models. Among these methods, models for cognitive diagnosis (CDMs) stand out because of their integration of fine-grained information on skills measured by the test within a psychometric framework. In this study, several IRT models and CDMs were fitted to the 4th grade TIMSS 2007 data (Mullis et al., 2007; Mullis et al., 2009) from three countries, which presented different average levels of achievement on the test. All models were compared in terms of absolute and relative fit, as well as with respect to item and person fit. Results favored the 3-PL IRT model in terms of relative fit, followed by the Additive CDM (de la Torre, 2011). Examinee A-CDM skill profiles differed by country, with the highest average level of achievement being the country with the highest proportion of examinees with mastered skills. In addition, latent skill estimates showed consistent positive correlations with the 3-PL IRT ability estimates, providing convergent evidence for the use of different psychometric models in the context of data from international educational assessments.

15 Multiple-group propensity score inverse weights truncation: Impact in bias reduction

Diego Luna-Bazaldua, *National Autonomous University of Mexico, Mexico*
Laura M. O'Dwyer, *Boston College*

Propensity scores methods have become a key technique to reduce bias in the estimation of causal effects in cuasi-experimental and observational research. Recent developments facilitate the estimation of propensity scores for treatment effects when more than two groups are compared. Previous studies have explored the role of trimming propensity scores inverse weights in designs with only one treatment and one comparison group, but no research-to-date has explored the implications of trimming inverse weights in the context multiple-group propensity scores. The present study adds to research on propensity scores by exploring the impact of truncating multiple-group inverse weights on covariate bias and recovery of the treatment effect parameter in a Monte Carlo study. The study consisted of four different conditions, which were defined based on the model used to generate the propensity scores and the treatment effect. Multiple-group propensity scores were estimated using three different models: multinomial logistic regression, generalized boosted models, and neural networks. Results from the simulation study indicate that trimming inverse weights increased covariate bias and did not have a substantial impact on parameter recovery statistics of the treatment effect. In addition, data mining methods tended to produce less extreme multiple-group inverse weights compared to multinomial logistic regression models.

16 Effects of random responses on latent class analysis

Yi-Kai Chen, *National Taiwan University, Taiwan*
Tong-Rong Yang, *National Taiwan University*
Li-Jen Weng, *National Taiwan University*

Information criteria (IC) are frequently used for deciding the number of classes in latent class analysis (LCA). The present simulation explored the possible effects of participant random responses on the performance of various ICs. Random responses constitute one form of aberrant responses that may affect the results of statistical analysis (e.g., Oshima, 1994; van den Wittenboer, Hox, & de Leeuw, 1997). In addition to the proportion of participants responding randomly, the number of classes of equal proportion, the number of variables, the size of conditional probabilities, and the sample size were also manipulated with 200 replications conducted for each condition. Seven ICs were evaluated including AIC, CAIC, sample size adjusted CAIC (CIAC*), AIC3, BIC, sample size adjusted BIC (BIC*), and HBIC. The behaviors of ICs varied greatly with number of classes. For the two-class models, all the ICs, except AIC and BIC*, selected the designated number of classes even when the random responses constituted as high as 20% of the sample data.

In contrast, for the four-class models, the probability of selecting the designated number of classes tended to decrease as the proportion of random responses rose especially at the case of weaker conditional probabilities where the likelihood of identifying one more class than the designated one increased. In view of the preliminary results from this exploratory study, the effects of random responses on the selection of number of classes and the quality of parameter estimates in LCA warrant further research attention.

17 An empirical examination of wording effects: An integrative approach

Fernando Ponce, *Pontificia Universidad Católica de Chile, Chile*
Álvaro Vergés, *Pontificia Universidad Católica de Chile*

This study aimed to evaluate empirical evidence of response style perspective of wording effects, using a person-centered statistical approach to a) identify a latent sub-group of individuals who exhibit a biased response pattern, based on the direction of the item wording (a different sub-group from those who exhibit, for example, high levels of the specific attribute), and b) use this classification to evaluate the presence of a differential pattern of factor scores in a bifactor solution, traditionally used to model method factors. The analysis consisted in Latent Profile Analysis applied to the Rosenberg Self-Esteem Scale data from the LISS Panel Longitudinal Study (n=6777). After recoding, the analyses were conducted using Mplus version 7.4. and maximum likelihood robust estimation. A 4-profiles solution was identified with optimal membership probabilities. This solution identifies a specific profile (8.6%) with an asymmetrical response pattern according to item-wording, and three substantive profiles (high, mid and low self-esteem). Finally, we used this classification to interpret the factor scores obtained from a bifactor solution composed by a general factor of self-esteem (SE) and a specific factor associated with negative wording (NW), commonly used to model wording effects. We found differential means in the latent variables factor scores in this specific biased profile, compared with substantive profiles. These results provide evidence for the notion that wording effects are not necessarily an inherent feature of a scale. Instead, they seem to emerge from the responses provided by a specific sub-group of individuals who can be characterized as biased responders.

18 Topic modeling of constructed-response answers on social studies assessments

Jiawei Xiong, *University of Georgia, United States*
Allan S. Cohen, *University of Georgia*
Hye-Jeong Choi, *University of Georgia*
Minho Kwak, *University of Georgia*
Seohyun Kim, *University of Georgia*

Recent qualitative evidence has suggested there is useful information in the text of answers to constructed response (CR) items (Buxton et al., 2014). Kim et al. (2017) found that results from a qualitative analysis of the text from student's CR answers were in close agreement with results from a statistical analysis of different samples from this same assessment. The statistical method used, Latent Dirichlet Allocation (Blei et al., 2003), employs Dirichlet distributions to model the number of latent topics and the number of words in each topic in textual data. LDA is referred to as an unsupervised method because it analyzes the text without considering other information. Blei and Mcauliffe (2008) developed the supervised LDA (sLDA) that combines the LDA with a label such as a score on an essay test. Results from Kim et al. (2017) and Kwak et al. (2017) suggest that the text in CR items is more highly constrained than in the kinds of documents for which topic models were initially developed. Consequently, the number of latent topics in a set of responses to CR items tends to be smaller (from 2 to 6), compared to what is extracted from less constrained text. In this study, we investigate the utility of the LDA and sLDA, to detect the latent thematic structure in a set of CR social studies assessments. We present an empirical example and a simulation study examining the effects of different sample sizes and priors on estimation of the two models.

19 In gathering evidence for Measurement Invariance: A Bayes Factor approach

Rui Jiang, *University of California, Davis, United States*
Philippe Rast, *University of California, Davis*

Measurement invariance (MI) is key to psychological assessment across groups or measurement occasions (Millsap et al., 2007). Establishing measurement invariance ensures the comparability of factor scores across conditions so that they can be meaningfully evaluated and interpreted (French et al., 2008). The majority of methods for assessing MI are based on multigroup CFA model comparisons via goodness of fit (e.g., ΔCFI, RMSEA) or χ2-difference tests. Typically, MI is assumed by a non-significant difference in these criteria. Yet, while the absence of significance is taken as supportive evidence for invariance, this is not a valid inference (Yuan et al., 2016).

In order to test for actual invariance across groups, we present an approach using Bayes factors (BFs) to gather evidence for MI. To evaluate the relative probability of one hypothesis (H0: invariance) over another (H1: non-invariance), the Savage-Dickey density ratio is used to compute a BF for each single parameter of interest. Following the guideline of categorizing BFs (Harold, 1961), we use a BF01 of 3 as substantial evidence for invariance, and 1/3 for non-invariance. A simulation study using Bayesian SEM was conducted to assess the performance of BF in testing MI. The simulation condition includes sample size, parameter difference, scale length, and item reliability. The results show that a BF approach performs well in supporting invariance and detecting non-invariance except when both the sample size and amount of non-invariance are small. We end by discussing implications for research on MI and more generally on equivalence testing.

20 Identifying referent variable(s) using constant anchor method in MIMIC-interaction modeling

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

Previous research has found that the all-other anchor method generally outperformed the constant anchor method for identifying a single referent variable. A possible explanation was that in the constant anchor method, chi-square difference statistics were erroneously inflated because each noninvariant observed variable was once used as an anchor variable to evaluate invariant observed variables. This study aims at demonstrating the advantage of designating a pair of referent variables over a single referent variable using the constant anchor method in the MIMIC-interaction model specification. A Monte Carlo simulation design was carried out to determine the effects of different number/percentage of noninvariant variables, magnitude of noninvariance, magnitude of group differences in latent means and variances, and sample size in a one-dimension measurement model with six continuous observed variables. Data generation and analysis were performed with Mplus 8. Accuracy rates for identifying referent variable(s) were evaluated by comparison with a benchmark criterion, the probability of randomly selecting truly invariant variable(s) from among all the observed variables. Results showed that for identifying one single referent variable, the constant anchor method performed fairly worse when (a) 50% of observed variables were non-invariant and (b) latent mean and variance differences were medium and large across groups in the conditions of 33% noninvariance. However, the constant anchor method performed well to locate a pair of referent variables across all conditions, even in the most extreme condition where 50% of observed variables were noninvariant and latent mean and variance differences were large across two groups.

21 The effects of DIF on IRT vertical scale

Hyesung Shin, *Yonsei University, South Korea*
Guemin Lee, *Yonsei University*

A vertical scale refers to a common scale applied to several grade levels for measuring student growth (Kolen & Brennan, 2014). To construct a vertical scale, the common item with non-equivalent groups (CINEG) design has been commonly used by most testing companies or institutes. For the CINEG design, the psychometric properties of common items would be crucial to a sound vertical scale. If common items in adjacent grade levels includes some differentially functioning items (usually called Differential Item Functioning: DIF), it might lead to a distorted vertical scale and biased estimate of student growth. Several studies examined the effects of DIF in common items on IRT horizontal equating (Huggins, 2014; Kim & Cohen, 1992; Candell & Drasgow, 1988). There has been little literature to investigate the effects of DIF in vertical scaling. This study was designed to investigate the effects of DIF in common items in achieving a vertical scale with a CINEG design. This study applies simulation techniques, with 18 conditions of 3 fully crossed factors: DIF directions (3 conditions), the number of DIF's (2 conditions), and DIF magnitudes (3 conditions). A parameter set to simulate the data without DIF can be served as a baseline to compare the results from various different DIF conditions. Preliminary analyses were completed and we found some relationship between conditions of DIF and biases of the vertical scale. The remaining analyses will be completed before the IMPS conference.

22 Designing an effect size to standardize polytomous item DIF

James Weese, *University of Arkansas, United States*
Ronna C. Turner, *University of Arkansas*

Classification guidelines of differential item functioning (DIF) in dichotomous items have been used (Zwick, 2012; Shealy & Stout, 1993), however guidelines for polytomous items have not been consistently agreed upon. This is partly due to the difficulty of making suggested guidelines for items of varying response ranges using methods that do not standardize variability. Zwick, Thayer, and

Mazzeo (1997) suggested creating an effect size for measuring DIF magnitude using SIBTEST. This would eliminate the need for varying criteria depending upon item response variation or other sample-specific factors. In this study, an effect size has been developed to be used with POLYSIBTEST for polytomous (and dichotomous) items that corresponds with the ETS and beta-uni dichotomous item classifications of DIF magnitude. An effect size is calculated using the beta-uni value that is a measure of unstandardized DIF magnitude. Dividing beta-uni by a group-weighted standard deviation creates a standardized effect size value that can be applied to dichotomous and polytomous item. The group-weighted standard deviation is calculated using the standard error of the test statistic and information about participant inclusion at each score level. A simulation study was conducted to first identify the effect size values that correspond with the ETS and beta-uni DIF magnitudes for dichotomous items. The selected dichotomous effect size value was then applied to polytomous data to demonstrate the relationship between prior DIF magnitude values that are associated with the effect size for items with ranging response options. Recommended effect size values of 0.14 and 0.20 will be presented.

23 On mean eigenvalues from random correlation matrices in parallel analysis

You-Lin Chen, *National Taiwan University, Taiwan*
Li-Jen Weng, *National Taiwan University*

Incorrectly choosing the number of factors could seriously distort the results of factor analysis. Parallel analysis proposed by Horn (1965) is one of the recommended methods for deciding the number of factors. Horn suggested comparing the eigenvalues of the data correlation matrix to the mean eigenvalues from several random correlation matrices to mimic the effect of sampling error, and hoped the theoretical distribution of the random eigenvalues could be derived in the future. This distribution was at last derived by Jiang in 2004 to follow the Marčenko-Pastur (1967) law at large sample sizes and large numbers of variables with multi-normal distribution by extending Johnstone's (2001) work from random covariance matrix to random correlation matrix. The behavior of mean random eigenvalues to their theoretical distribution deserves research attention, especially under finite sample sizes and limited numbers of variables usually encountered in psychological researches. This study compared the mean eigenvalues from 100, 500, or 1000 random correlation matrices to their theoretical counterparts at various sample sizes (50, 100, 200, 300, 500, and 1000) and number of variables (5, 10, 15, 20, 30, 40, and 50). The discrepancy between empirical and corresponding theoretical eigenvalues was found to be about .04 or less across all the simulated conditions. The results support Horn's idea of using mean eigenvalues from a set of random correlation matrices to approximate the theoretical distribution of the eigenvalues of correlation matrix with uncorrelated variables taking sampling error into account.

24 The Schmid-Leiman solution in ETA psychometric validation

Cecilia Zaidán, *Universidad de la República Uruguay, Uruguay*
Victor E.C Ortuño, *Universidad de la República Uruguay*

Hierarchical solutions for Exploratory Factor Analysis are varied and little known. In the case of this research, the Schmid-Leiman Solution was used to test the factorial structure of the Triangular Love Scale (ETA). This questionnaire aims to evaluate the level of relationship between couple relationships through three main components; Intimacy, Passion and Commitment. At the psychometric level, within Latin America, the ETA presented inconsistencies. After a thorough investigation, we worked to improve the model with the Schmid-Leiman structure. The procedure allows us to test a second order hierarchical model where the first order factors are orthogonalized among themselves allowing us to interpret the relative impact of one or several general second order factors (Wolff & Preising, 2005). The results obtained improve the performance of the scale at the psychometric level. A general factor of second order was found plus three factors that respond to the theory that gives rise to the evaluation. Results of alternative structural models are presented as well as a reduced version of the inventory.

25 Response styles analysis with different approaches

Sohee Kim, *Oklahoma State University, United States*
Hyejin Shim, *University of Missouri*
Ki Lynn Cole, *Oklahoma State University*

When people respond any survey, their answers to the items to which they respond are valuable. However, those responses could be influenced by other factors, such as the response rating scale. This non-content-based form of responding are usually referred to as response styles. Response styles can be a source of contamination in self-report survey that are used widely in the educational and psychological researches, and therefore they threaten the validity of conclusions drawn from research data. The

purpose of current study is to analyze and compare three different approaches for assessment and control for response styles by utilizing a mixed Rasch model (Rost, 1990), multidimensional partial credit model (Kelderman, 1996) and IR tree model (Bockenholt, 2012; DeBoeck & Partchev, 2012). Responses to the PANAS scale was used, and data were analyzed with Mplus (Muthén & Muthén, 2007). To evaluate different models for response styles, the values of AIC and BIC were compared and the estimated latent traits also were compared. According to the results of AIC and BIC, all multidimensional PCMs fit better than other models. For the PANAS scale, the best-fitting model was the multidimensional PCM with the dimensions PANAS and DRS. The importance of considering response styles based on the self-efficacy questionnaires was explored in the current study by comparing different approaches. Moreover, it allows more flexibility of latent variables for analyzing an item response process by utilizing a tree structure or multidimensional approaches, instead of only focusing on the item responses.

26 Performance of S-X^2 item fit index for polytomous IRT models

Moo Won Kwon, *Yonsei University, South Korea*
Guemin Lee, *Yonsei University*

Orlando and Thissen's S-X^2 item fit index has performed better than traditional item fit statistics such as Yen's Q_1 and McKinley and Mill's G^2 in applying item response theory (IRT) models to tests with dichotomous item (Orlando & Thissen, 2000, 2003). The utility of Orlando and Thissen's S-X^2 item fit index was extended to tests with polytomous items by Kang & Chen (2008, 2011), including the generalized partial credit model (GPCM), partial credit model (PCM), and graded response model (GRM). Previous studies investigated the performance of the generalized S-X^2 item fit index with limited conditions of simulation such as fixed number of response categories and fixed test lengths. This study was designed to investigate the performance of the generalized S-X^2 with various conditions encountered in testing practices. Eighteen conditions [(3 sample sizes) $\times$ (2 response categories) $\times$ (3 test lengths)] were considered in this study to examine the performance of S-X^2 item fit index. Type I error rates and powers were examined to evaluate S-X^2 item fit index with three polytomous IRT models (GPCM, PCM, and GRM). Each polytomous IRT model was used as data generation model (GM) and calibration model (CM), respectively. The performance of S-X^2 item fit index was examined when the GM and the CM were identical and different.

27 Generalizability theory approaches in estimating conditional SEM of testlet-based tests

Heuijo Lee, *Yonsei University, South Korea*
Guemin Lee, *Yonsei University*

Using generalizability theory three models were compared when estimating conditional standard errors of measurement (SEM) of test scores composed of testlets having random effects. When several items are grouped into testlets, especially reading comprehension passages, the assumption of the conditional independence among items could be violated. This study was aimed for investigating how the estimates of conditional SEM of testlet-based test scores were influenced as the testlet effects and the degrees of imbalance of items across testlets increased. Three generalizability theory models were compared: item score model (p$\times$I), testlet score model (p$\times$H), item nested testlet model (p$\times$(I:H)). In the item score model, each item was treated as an independent item; in the testlet score model the item scores sharing common testlets were summed up and treated as a single score at each testlet; in the item nested testlet model the testlet structure was reflected. Based upon results from simulated data, the conditional SEM seemed underestimated with the item score model and the magnitude of the negative bias increased as the testlet effects and the degree of imbalance increased.

28 A new family of unidimensional item response models with asymmetric heavy tailed distributions and item response functions

Caio Azevedo, *State University of Campinas, Brazil*
Juan. L. Padilla, *Intelligent Process Enterprise*

Item response theory (IRT) comprises a set of statistical models that are useful in many fields, as psychometrics, in which there is an interest in measuring the so-called latent traits. Such models are used to quantify the probability of subjects get a certain score in items which belong to a certain measurement instrument, as a function of those latent traits and item parameters. Two assumptions usually considered in many IRT models are: to assume that the latent traits follow symmetric normal distributions and to use symmetric regular tails link functions, in particular the probit and logit links, as the Item Response Function (IRF). However, there are examples in the literature where the data sets do not support such assumptions and they are only used for convenience. Furthermore, it has been evidenced that model misspecification, in both latent traits distribution and IRF, can lead to biased estimates and misleading conclusions. The aim

of this work is to propose a flexible family of unidimensional multiple group IRT (UIRT) models for dichotomous data, allowing for asymmetric/heavy tailed components in both latent traits and IRF. Also, it wasevidenced that our proposed UIRT models perform better under the presence of heavy tails, in the two components. We focus on the centered skew-t distribution, to define the IRF and the distribution of the latent traits. A full Bayesian approach for parameter estimation and model fit assessment is developed, through suitable MCMC algorithms. Results of simulation studies indicate that the model/estimation methods recovery the parameters properly and that the model fit and comparison tools behaved well. In addition, we analyse a real data set corresponding to a large-scale educational assessment, where it is shown that our approach performs better than the usual ones.

29 Uncovering multiple strategies in an exploratory cognitive diagnostic modeling framework

James Balamuta, *University of Illinois at Urbana-Champaign, United States*
Steven Culpepper, *University of Illinois at Urbana-Champaign, IL, USA*
Jeff Douglas, *University of Illinois at Urbana-Champaign, IL, USA*

The development of exploratory cognitive diagnostic models (ECDMs) broadens the applicability of diagnostic models and provides a framework for evaluating expert pre-specified Q-matrices. Previous research on exploratory diagnostic models assumed respondents use a single strategy or set of attributes when solving a problem. We develop new Bayesian methods for estimating multiple strategies of problem solving under the exploratory framework. We provide Monte Carlo simulation studies showing parameter recovery. Furthermore, we apply the method to a spatial rotation dataset.

30 Some new scale transformation methods

Kuo-Feng Chang, *University of Iowa, United States*
Won-Chan Lee, *University of Iowa*

Scale transformation methods are important psychometric tools to develop a common metric so that examinee performance across different test administrations are comparable. By far the two most common scale transformation methods in operational tests are the test characteristic curve (TCC) methods by Haebra (1980) and Stocking-Lord (1983) by using the 3PL model (Birnbaum, 1968). This study aims to propose and examine two types of modified methods including 3PLab TCC method (by ignoring c parameter estimates and only using a and b parameter estimates in the TCC methods) and robust item characteristic curve (ICC) methods (by first conducting single-item linking using each linking item and then applying weights to compute final linking coefficients) by using Monte-Carlo simulations. Three factors are manipulated in this study: (a) the number of linking items-10, 20, and 30; (b) sample size-1000 and 3000; and (c) group difference-N(0.5,1) and N(1,1) (reference group is from N(0,1) in this study). Evaluation criteria include: conditional and overall bias, and conditional and overall root mean squared error. The results from the original Stocking-Lord (SL) and Haebra (HB) methods are used as a benchmark. The preliminary results showed that robust ICC tended to produce better recovery results for b parameter among various methods. In addition, 3PLab of SL method generally outperformed the original SL method across various conditions, and 3PLab of HB method tended to produce better recovery results than the original method under the conditions with a small sample size. The current study results will be helpful for measurement practitioners.

31 Extending the K-index for answer copying detection

Shu-Ying Chen, *National Chung Cheng University, Taiwan*
Hsiu-Yi Chao, *National Taiwan University*
Jyun-Hong Chen, *Soochow University*
Chuan-Ju Lin, *National University of Tainan*

Among several methods proposed for answer copying detection, the K-index (Holland, 1996) is adopted by test practitioners and has been proven its capability. However, it has several limitations that restrict its practical applicability, including the lack of an objective method for estimating the Kling function. To improve the applicability and objectivity of the K-index, the authors (2017) have proposed the KE-index with exact probability model for answer copying detection and proved its efficacy. In this study, the KE-index was further extended with considerations of the matching variable and the matching responses to improve its efficacy in detecting answer copying. A series of simulation studies were conducted to examine the performances of the KE-index with two matching variables (number correct score / trait estimate) and two matching responses (incorrect / both correct and incorrect responses). The results indicated that Type I error rates were well-controlled for all the indexes when the estimation error was not large. When the difference between source's and copier's abilities was small, the KE-indexes with incorrect responses as matching responses

generally outperformed those with all responses as matching responses. For conditions with large difference between source's and copier's abilities, the KE-index with all responses as the matching responses and trait estimate as the matching variable outperformed the other methods in power rates. Based on the results, the KE-index shows promise for answer copying detection in practical applications.

32 Item response theory using autoencoders and variational autoencoders

Claudia Evelyn Escobar Montecino, *Universidade Federal de São Carlos , Brazil*
Mariana Curi, *Universidade de São Paulo*

Due to the development of technology, pattern recognition has evolved into multiple methods that help computers make predictions and make larger database decisions. One of the goals of these methods is to reduce the size of input data, which is interesting for Item Response Theory (TRI) that attempts to quantify an individual's latent abilities from his/her item responses in a test. With this in mind we consider a Multidimensional Logistic Model of two Parameters (MP2L) to simulate the responses of a group of individuals in a test that evaluates certain abilities, then we estimate these competences through two variations of neural networks: Autoencoders e Variational Autoencoders, and we use metrics to compare them with each other and to the usual estimation techniques, like MCMC and maximum likelihood. The methods are also applied to a real database referring to a test that follows the pattern of the simulated data in the previous step. These methods are being proposed as an alternative to the commonly used estimation in TRI.

33 Normativity of trait estimates from multidimensional forced-choice data – A simulation

Susanne Frick, *University of Mannheim, Germany*
Anna Brown, *University of Kent*
Eunike Wetzel, *University of Vienna*

The Thurstonian item response model (TIRT) allows deriving normative trait estimates from multidimensional forced-choice (MFC) data. In the MFC format, persons must rank-order items that measure different attributes according to how well the items describe them. The aims of this study were to compare trait estimates from the TIRT to trait estimates derived from classical (partially) ipsative scoring, from dichotomous true-false data, and from rating scale data, and to evaluate normativity with unbiased item parameters and trait correlations. For a questionnaire with five traits, MFC data were simulated for blocks of three items and rating scale data were simulated for a five-point scale. The data were analyzed classically (mean scores) and with the TIRT, normal ogive, or graded response model, respectively. Additionally, four factors were varied and completely crossed: trait intercorrelations, item keying, questionnaire length, and number of items per trait. To evaluate trait estimation, trait recovery, reliability, and several bias measures were computed. The results showed that TIRT trait estimates are normative in contrast to partially ipsative ones. However, longer questionnaires or larger blocks are needed to achieve the measurement precision of rating scale data. Thus, TIRT modeling of MFC data overcomes the drawbacks of the classical scoring approach.

34 An IRTree approach investigating the construct dependency of response styles

Nikole Gregg, *James Madison University, United States*
Brian C. Leventhal, *James Madison University*

Variability in survey responses is not only due to the respondent's attitude of interest but may also reflect an individual's response style. Without considering response styles, such as extreme response style (ERS), researchers may inaccurately interpret respondent attitudes. Thus, researchers have proposed models, such as the IRTree model, to account for and investigate ERS (e.g., Thissen-Roe & Thissen, 2013). Extreme responding is often correlated with several traits such as self-concept clarity and simplistic thinking (Cabooter, et al., 2016; Naemi, Beal, & Payne, 2009). Furthermore, current research suggests ERS may be domain specific (Cabooter, Weijters, De Beuckelaer, & Davidov, 2016). In this study, we investigate whether ERS may be domain specific using Likert-type items designed to mimic a two-stage sequential IR-Tree model. Six hundred undergraduate students from a mid-size university responded to item prompts from Task Value and Self-Efficacy subscales. These prompts were presented in two-stages, where each stage corresponded to a node on an IRTree model. During the first-stage, respondents decided whether they disagreed, were neutral to, or agreed with an item prompt. Response options at the second-stage were conditional upon the respondents' first choice. For example, if a respondent chose 'Disagree' at the first-stage, they then reported the intensity of that response by selecting between 'Strongly Disagree' and 'Disagree'. Using this two-stage item presentation, we found that the frequency of extreme response selections differed compared to traditional Likert-type items.

Using node-level data as compared to traditional item-level data, we estimate IRTree models to investigate the construct-dependency of the ERS.

35 Differential validity of cognitive models for concept learning among latent classes

Clifford Hauenstein, *Georgia Institute of Technology, United States*
Susan Embretson, *Georgia Institute of Technology*

The Concept Formation subtest of the Woodcock Johnson III (WJ III) requires examinees to employ deductive reasoning and rule-based categorization skills. Each item requires examinees to derive a particular rule that identifies the geometric commonalities among a set of presented shapes. Using responses from the 2001 WJ III norming sample, the authors previously found that examinees demonstrated greater struggle with items involving disjunctive rules over conjunctive rules. This is consistent with a literature base pointing towards a natural preference for conjunctive rules. However, more contemporary efforts in modeling concept learning processes have focused on representing these tasks with Boolean functions. Feldman (2000) claimed a predominance of the essential dimensionality of the rule in determining task difficulty. Feldman updated his approach by evaluating the underlying algebraic complexity of concepts. More recently, Goodwin and Johnson-Laird (2011) considered the relationship between working memory burden and Boolean structure of concepts. We ask whether distinct classes of individuals can be identified for which one approach is most relevant in describing the process of rule derivation. If so, it may indicate differences in invoked strategy across examinees, and help address the differential validity of concept learning models. Von Davier and Rost's (2016) mixture Rasch model is applied to identify discrete classes of examinees for whom different patterns of item difficulty parameters might be observed. To explore the meaningfulness of the resultant classes in accordance with difficulty models from Feldman and Goodwin and Johnson-Laird, Fischer's linear logistic test model (1973) is applied to each group separately.

36 Bi-Factor model of CASP-12 for general QoL factor in older adults

Matthew Kerry, *Zürich University of Applied Sciences , Switzerland*

Patients' subscores on quality of life (QoL) measures can provide diagnostic information about strengths and weaknesses of respondents' performance in specific areas. Such diagnostics may help with identification of potential at-risk individuals. Subscores may also help with modifying extant care-treatment programs. The Control, Autonomy, Self-realization, and Pleasure (CASP) measure is one, popular QoL measure example with such subscore potential, which will be of focal interest and presented via the current poster submission. As the CASP's author reassures researchers that "those who simply require a single index" may sum the CASP-12, it is important to first-determine if unidimensional usage in prediction models is reasonably unbiased by ignoring subdomains. As the CASP constructor's concluded, "... strength of the inter-domain correlations....confirm our belief that QoL is a unitary phenomenon which is the product of the interactions between the domains". In this first-IRT inspection of CASP's psychometric properties, the CASP-12's general QoL factor was found to be well-specified by a bi-factor model for specifying subdomains/content homogeneity as sources of nuisance variance. Furthermore, the CASP-12's total score (general factor) exhibited acceptably high reliability in older populations across both broader community-dwellers, as well as among narrower-patient respondents. In contrast, the CASP-12's specific subfactors were found to exhibit unacceptably low reliability, suggesting only CASP-12's global score is currently appropriate for substantive interpretation and meaningful use. Finally, the CASP's original 12-item measure was identified as-having a potentially useful, 5-item subset for succinct indexing of QoL-unitary scores for future researchers' use in structural-estimation models.

37 IRT Assessment of Readiness for Interprofessional Learning Scale(RIPLS): Dimensionality, reliability, and item function

Matthew Kerry, *Zürich University of Applied Sciences, Switzerland*

This poster presentation aims to bring evidence from modern psychometric methods to bear on a popularly deployed questionnaire in interprofessional education (IPE) assessment. Specifically, three interrelated problems raised against the Readiness for Interprofessional Learning Scale (RIPLS) are examined in a study with n =280 medical and nursing student participants. Firstly, findings indicate a strong, general factor underlying the RIPLS that supports unidimensional interpretations. Secondly, findings support RIPLS overall reliability, but fail to support subscale reliabilities. Thirdly, findings support the RIPLS potential sensitivity to changes with appropriate lower ranges for our pre-training student sample. Recommendations for refinement to the RIPLS include: use of

more appropriate reliability indices; factor generalizability; and a subset of items. More generally, refinement is possible, whereas RIPLS disuse or continued misuse with problematic scales is likely to hinder progress in the field of IPE research.

38 IRT methods estimating CSEM for testlet with imbalance

Haejin Kim, *Yonsei University, South Korea*
Guemin Lee, *Yonsei University*

Testlet is defined as subgroup of items, sharing the same stimulus. Most achievement and aptitude tests were composed of passages or contents associated subgroups of items regarded as testlets. The number of items per testlet are different across testlets, which is called unbalanced structure. Despite of the imbalance structure, most previous studies investigated testlet effects for the balanced situation. The purpose of this study was to investigate the relative appropriateness item response theory(IRT) method for testlets composed of unbalanced design. This study was designed to compare two IRT methods(two parameter logistic model, and graded response model), for estimating conditional standard errors of measurements(CSEM). The simulated data was composed of testlet models varying degrees of imbalance and different testlet effect. Two IRT models are the two parameter logistic (2PL) model and graded response model(GRM). The former is item-based method, and the latter is testlet-based method. The results of this study were analogous to the results of previous studies in estimating the CSEM of a test composed of tesltets. Both balanced and unbalanced tests showed similar pattern of underestimation, and magnitudes of underestimation. Generally, 2PL model underestimated the CSEM rather than GRM, and the magnitude of underestimation using 2PL model was affected more with testlet effect than degrees of imbalance of testlet items.

39 Four MIRT equatings of testlet-based test scores

Guemin Lee, *Yonsei University, South Korea*
Hyejin Kang, *Yonsei University*

Testlets, as the name implies, have been defined as smaller subsets of a larger test (Wainer & Kiely, 1987). Conventional unidimensional item response models may not entirely reflect the underlying data structure of testlet-based tests. One way to take testlet effects into account can be to incorporate more dimensions or factors in addition to a general dimension. The residual dependence within testlets after controlling for the influence of the primary factor could be modeled by introducing additional latent dimensions or factors. Lee, Kolen, Frisbie, and Ankenmann (2001) was probably the first that investigated the effects of testlets in the context of IRT equating. Li, Bolt, and Fu (2005) provided a test characteristic curve linking method under the testlet response model. Recently, Lee, et al. (2015) provided a bi-factor MIRT true-score equating for testlet-based tests and Lee and Lee (2016) developed a bi-factor MIRT observed-score equating for mixed-format tests. Tao and Cao (2016) examined an extension of IRT-based equating to dichotomous testlet response theory model. Lee, et al. (2017) explored the possibility of higher-order IRT model to equating for testlet-based test scores. The main purposes of this study are to specify four different MIRT models, (a) testlet response model, (b) bi-factor model, (c) simple structure model, and (d) second-order model, for testlet-based tests and to evaluate relative appropriateness of those models in the context of equating. Equating methods using unidimensional dichotomous IRT (2PL) and conventional equipercentile procedures are also implemented for the purpose of comparison with specified models in this study.

40 Estimating racial bias in trauma screening using IRT methods.

Isis Martel, *University of Arkansas, United States*
Latunja Sockwell, *University of Arkansas for Medical Sciences*

This study was conducted to pilot a strategy for better understanding differences in trauma between Black/African-American and Caucasian adults undergoing substance abuse treatment. Analyses were conducted to determine which proportion of group differences on trauma symptoms were due to bias via differential item functioning (DIF) and/or true latent trait (trauma) differences within this sample. Participants were 255 adults aged 18 and older participating in inpatient substance abuse treatment in Arkansas, USA. The trauma screening used in this study is the Civilian version of Post Traumatic Checklist -17. The analytic strategy was a graded response model of item response theory. Most items displayed DIF when grouping by race (African American versus Caucasian), accounting for most of the group difference. After controlling for DIF, the group difference that remained could be attributed to covariates such as childhood experiences, SES, history of substance use, legal history, and parent/family demographics. These pilot findings point toward the need for the development of culturally competent measurement tools specifically for historically underrepresented and vulnerable populations, and replication within a larger sample size to increase power.

41 Psychometric properties of irrational procrastination scale using item response theory

Shirin Rezvanifar, *Allameh Tabataba'i University, Iran*
Ali Delavar, *Allameh Tabataba'i University*
Hassan Mahmoudian, *Allameh Tabataba'i University*

In procrastination, people tend to avoid an activity and postpone it to future. Different tools have been developed to measure procrastination. The purpose of this study was to evaluate the psychometric properties of irrational procrastination scale (IPS) (Steel, 2002) items using item response theory (IRT). A total of 414 (50.2% females and 49.3% males) Iranian students participated in this research. Its factor structure was compared in replication analyses, and threshold parameters of item difficulty and discrimination power were estimated for each item using IRT. The data were analyzed via Graded Response Model (GRM). The function information of the questions showed that items two and nine represented little information. The scale showed good psychometric properties and appears to be an appropriate measure of procrastination attributes.

42 An IRT model for zero-inflated data

Hyejin Shim, *University of Missouri - Columbia, United States*
Wes Bonifay, *University of Missouri-Columbia*

Measurement researchers and practitioners often struggle with the question of selecting the most appropriate item response theory (IRT) model for their particular test data. The most widely applied unidimensional IRT models assume that the latent variable underlying the observed responses is normally distributed and that the item characteristic curves are symmetrical around an inflection point. Although these assumptions are statistically and interpretatively convenient, they are untenable in the presence of zero-inflated data. Typically, violations of these assumptions are handled through carefully specifying the prior distribution(s), fitting complex non-parametric or non-normal item response functions, or simply throwing out the problematic data patterns and thereby discarding potentially useful information. To more appropriately analyze zero-inflated data, we introduce an asymmetric dichotomous IRT model based on the complementary log-log (CLL) link function. In addition to allowing for irregular response probability functions, the 1-parameter CLL model is potentially more parsimonious than alternative IRT models. In this research, we use simulation and empirical results to establish the statistical accuracy and interpretation of the CLL model relative to more traditional IRT models. Specifically, we demonstrate that the CLL model provides better goodness-of-fit than the 1-parameter logistic model when the population distribution is skewed toward the low end of the latent trait scale. Based on our findings, we conclude that application of the CLL link function in IRT modeling will yield reliable and interpretable parameter estimates in the presence of many zeros, with no need for complex modeling approaches or unnecessary data transformations.

43 Comparing flexMIRT and the R package 'mirt' for MIRT models

Kun Su, *University of North Carolina at Greensboro, United States*

In this study, the author investigated the model parameter recovery of two software packages, flexMIRT and R package 'mirt' via a series of simulation conditions. Simple structure, approximately simple structure and bifactor models are examined in the study by two item parameter estimation techniques, the Metropolis-Hastings Robbins-Monro (MH-RM) algorithms, and the Marginal Maximum Likelihood Estimation (MMLE). This is an ongoing research, and the preliminary results showed that the two software programs had similar root mean square error values. Overall, the flexMIRT was found to have smaller RMSE and the R package 'mirt' was found to have shorter estimation time.

44 An extended multi-process model for wording effects in mixed-format scales

Yi-Jhen Wu, *University of Bamberg , Germany*
Kuan-Yu Jin, *The University of Hong Kong*

Likert-scale items are widely implemented in psychological tests and social surveys. They are either positively or negatively worded. When all items are not written in the same wording direction, it would lead to unexpected nuisances so-called as "wording effects". How the wording effects influence the response processes remains unknown. Recently, several multiple-process IRT models, which are very flexible and helpful to account for individual response process, have been proposed to analyze Likert-scale items. Based on Böckenholt's three-decision model (2012, 2017), we propose a new model by incorporating additional random variables to quantify the wording effects in different response processes. The subscale of extraversion from Big Five personality inventory, including six positively-worded items and six negatively-worded items, was used to demonstrate. Because of the high dimensionality of the new model, the Marcov chain Monte Carlo method was adopted for parameter estimation in

WinBUGS. The results fit our expectation that ignoring the wording effects would overestimate test reliabilities of the three latent traits. Moreover, the preliminary findings suggested that there were no consistent patterns of wording effects for the negative items across the three processes. More real examples are needed to validate the conclusions.

45 Longitudinal associations between student engagement and their relationships with teachers and peers

Lina Geng, *Beijing Normal University, China*
Lingyan Li, *Beijing Normal University*

The purpose of the study was to examine the longitudinal and reciprocal relations between students' perceptions of teacher-student relationships and peer relationships and student engagement (behavioral, emotional and cognitive engagement) in China. Participants were 628 7th grade students from one public secondary school. All the students completed the self-reported student engagement, teacher-student relationships and peer relationships measures at three time points over a period of three years (end of seventh grade, end of eighth grade, and end of ninth grade). Through cross-lagged models, this study found that teacher-student relationships at the prior year could only predicted the cognitive component of student engagement at the end of ninth grade. But, higher levels of behavioral, emotional and cognitive engagement in earlier times predicted subsequent higher quality of teacher-student relationships across 7th to 9th grade. Antecedent peer relationships predicted student behavioral and cognitive engagement at eighth grade and emotional engagement at ninth grade. In turn, behavioral, emotional and cognitive engagement at seventh grade predicted peer relationships at eighth grade, but the prediction was only significant from cognitive engagement to peer relationships for older students between the end of eighth grade and end of ninth grade. Further, the effects of earlier peer relationships on cognitive engagement at 8th grade, and teacher-student relationships at 8th grade on subsequent emotional and cognitive engagement were more positive in boys than girls in secondary school.

46 Can we distinguish between longitudinal models for estimating nonlinear trajectories

Ai Ye, *University of North Carolina at Chapel Hill, United States*
Kenneth A. Bollen, *University of North Carolina at Chapel Hill*

Longitudinal data (aka. panel data, or repeated measures) are widely available in social science research. One of the challenges to analyze longitudinal data is that there usually is a lack of substantive theory to dictate a particular form of growth trend that can guide the selection of the optimal statistical model. Given its importance, the purpose of the present study is to investigate to what extent the optimal longitudinal model can be distinguished from alternative ones, with respect to goodness-of-fit statistics and parameter estimates, in the absent of theoretical hypotheses. The present work focuses on longitudinal data following an overall nonlinear trajectory, that is, a growth trend that cannot be represented by a constant rate of change. Because exploratory approach such as visual inspection of plots to search for the optimal model could be particularly problematic. The investigation includes a variety of common longitudinal models such as autoregressive models, linear and nonlinear latent growth curve models, linear and nonlinear autoregressive latent trajectory models. This study is performed via an empirical simulation design in which population parameter values of data generating models are carefully selected from previous studies published at peer-reviewed journals. We examine the extent to which the false models fit significantly worse than the true model, and compare the model-implied growth pattern discovered by each model. In addition, we observe other potential consequences by fitting the data by a false model with the wrong set of parameters. Lastly, I make recommendations for model selection and testing procedure.

47 Bi-factor MIRT observed-score equating for mixed-format tests with CINEG design

Seonghyun An, *Yonsei University, South Korea*
Guemin Lee, *Yonsei University*

Mixed-format tests, which contain both multiple-choice and free-response items, have been widely used in educational settings. Because different item types sometimes measure different constructs, mixed-format tests can introduce a multidimensionality. If the mixed-format tests would be multidimensional, applying conventional unidimensional IRT model might lead to biased results in the context of equating. There have been a few previous studies that applied MIRT models to mixed-format tests. Most of them focused on a random groups design. Lee and Lee (2016) proposed a procedure of Bi-Factor MIRT observed-score equating for mixed-format tests with random groups design. There would be some difficulties in using random groups design in practice because test forms to be equated should be administered

at the same time. A common-item nonequivalent groups (CINEG) design, on the other hand, can be alternative to random groups design in that the test forms could be administered at different times. Kim (2017) implemented CINEG design for MIRT observed-score equating by using concurrent calibration for scale linking procedure. However, this study contained only multiple-choice items. This study was designed to expand his procedures to mixed-format tests composed of both multiple-choice items and free-response items. The Bi-Factor MIRT equating results were compared with those from a unidimensional two-parameter logistic model across varying conditions: the multiple-choice and free-response item composition and proportion of common items relative to total items, and the correlations between dimensions for multiple-choice and free-response item type.

48 Dealing with item-level missingness in a multilevel data structure

Ye Feng, *Fordham University, United States*
Heining Cham, *Fordham University*

Missing data is pervasive in large-scale survey research with multiple scale measurements and nested data structures. While there are some suggestions on how to handle item-level missing data, there are no methods proposed and studied on how to handle these missing values in clustered data structures. I studied multiple imputation methods on item-level missing data using one real dataset. I compared the various methods of multiple imputation and studied the influences of features of complex surveys such as clustering, categorical variables.

49 Missing imputation for inflated count data

Ting Hsiang Lin, *National Taipei University, Taiwan*

Modeling count data is a topic of great interest in survey, psychology, medicine and other fields. For some questions, we often see a mass of some specific values, for example, zeros, and it is called inflated data. Another common problem in survey is non-response or missing data, and it often arises when we do not get adequately completed responses from the subjects. Non-response can occur due to two reasons: non-contact of selected subjects or the subjects' refusal to participate fully or artially. There are many situations for inflated data and missing data occur simultaneously. In terms of handling incomplete data, there are several methods including deleting, weighting and imputation. In this study, we would like to investigate the performance of different imputation methods when missing data is inflated and compare the performance of difference imputation methods. We will focus on model –based imputation with prior information and compare several inflated models. The following models are considered: generalized linear regression (GLM) with Poisson or with negative binomial distribution, zero-inflated regression with Poisson (ZIP) or with negative binomial distribution (ZINB), Zero-and K-inflated Poisson regression (ZKIP) and Multiple inflated Poisson regression (MIP). In this study, we studies the missing values of data with excess zeros. We investigated what are significant factors on the accuracy rate of imputation for inflated data with missing values. These factors include imputation methods, missing data mechanism, sample sizes, proportion of inflated value and missing rate.

50 Testing heterogeneity in inter-rater reliability estimation

František Bartoš, *Charles University, Czech Republic*
Patrícia Martinková, *Czech Academy of Sciences and Charles University*
Marek Brabec, *Czech Academy of Sciences*

Previous studies demonstrated that inter-rater reliability (IRR) may be affected by contextual factors such as age, gender, applicant's internal vs. external status or a field of study of a grant proposal. The heterogeneity in IRR can be tested and accounted for using different methods, including mixed-effect models, as in Martinková, Goldhaber & Erosheva (2018) or Generalized Estimating Equation (GEE) as in Mutz, Bornmann & Daniel (2012). This study presents a series of simulations comparing these approaches for testing heterogeneity in IRR. We consider several true data generating mechanisms to simulate datasets with continuous and categorical outcome variables with heterogeneity in IRR caused by either a continuous or categorical contextual variable. Subsequently, we fit Bayesian and non-Bayesian mixed-effect models as well as GEEs to estimate IRR. We use Likelihood ratio test, information criteria, and leave-one-out cross-validation, to select optimal model and to test for IRR moderators. Methods are compared in terms of type I. error rate and power.

51 A response time model to test with limited time

Sandra Flores Ari, *Sao Paulo University, Brazil*
Jorge Luis Bazán, *São Paulo University*
Heleno Bolfarine, *São Paulo University*

Response time (RT) on Assessment has been modeled using positive continuous distributions as lognormal, gamma, exponential and weibull distributions. However,

usually the time to complete a test is not infinite but it is limited because usually Test has a defined time to be completed. Thus, the purpose of this work is to propose a new RT model following a bounded distribution which can be more appropriate to data set that is obtained in Assessment with limited time to complete the Test. We consider a Bayesian approach. By considering simulation study we show that assumption of positive distributions to RT, when in fact the time of Test is limited, can be inadequated. Additionally, by considering data from PISA 2015 computer-based reading, we show the advantages of the new proposed model.

52 Investigating the effect of high school curriculum type with multilevel SEM methods

Burhanettin Ozdemir, *Siirt University, Turkey*

This study aims at investigating the effect of high school curriculum type favored by schools on students' nationwide general aptitude test (GAT) scores. For this purpose, multilevel-SEM methods were employed to examine the curriculum type effect on quantitative and verbal GAT scores along with other variables, such as gender at student-level, school size, school area and school type at school-level. Moreover, multilevel multi-group SEM method was employed to compare effect of these factors between public and private schools. The data set comprised of 29,203 high school students which was drawn from population with stratified cluster sampling method to ensure that nested structure data remained same. The model comparison statistics imply that random-intercept model showed better fit compared to other models and between level variability explained 17 to 25 percent of variance in total variance. These results indicate that both verbal and quantitative scores vary across schools which requires using multilevel SEM models. Multilevel multi-group SEM results reveal that gender effect on quantitative scores was not statistically significant while this effect was significant on verbal scores for each school type indicating that gender effect in favor of female students on verbal scores might be an indicator of either differential item functioning or true ability difference. Moreover, the effects of curriculum type variable on outcome variables were significant for both public and private school and, these effects were somewhat larger for private schools, which indicate that change in curriculum type (course-based curriculum) caused significant increment in verbal and quantitative scores.

53 On the precision matrix in high dimensional settings

Kentaro Hayashi, *University of Hawaii at Manoa, United States*
Ke-Hai Yuan, *University of Notre Dame*

Many aspects of multivariate analysis involve obtaining the precision matrix. When the dimension is larger than the sample size, the sample covariance matrix is no longer positive definite, and the inverse does not exists. Under the sparsity assumption, the problem can be dealt with by methods such as the graphical models and linear programming. However, in high-dimensional settings in behavioral sciences, the sparsity assumption does not necessarily hold. The dimensions are often greater than the sample sizes while they might still be comparable. Under such circumstances, introducing some covariance structures might solve the issue of estimating the precision matrix. We examine such a strategy and compare different approaches, and show that some approach gives relatively small mean square errors when the dimensions are larger than the sample size. However, if the dimensions are much greater than the sample sizes, the covariance structure approach may fail.

54 Application of multilevel redundancy analysis to hierarchically structured large-scale educational data

Hongwook Suh, *Nebraska Department of Education, United States*
Kwanghee Jung, *Texas Tech University*
Jaehoon Lee, *Texas Tech University*
Kyungtae Kim, *Tennessee Department of Education*
Jungkyu Park, *Kyungpook University*

Multilevel redundancy analysis (MLRA) can optimally handle hierarchically structured large-scale data with its dimension reduction feature. This study illustrates applications of MLRA for educational research, especially its use with state-wide standardized test data that often involve many levels as well as a large number of predictors and/or criterion variables. MLRA first decomposes variability in the criterion variables (e.g., subscores on a standardized test) into several orthogonal components corresponding to the predictor variables at different levels (e.g., student-level predictors and school-level predictors), and then applies singular-value decomposition to the decomposed parts to find more parsimonious representations (i.e., smaller number of principal components derived from the predictor variables). Thus, MLRA can evaluate a large-scale predictive model in a more parsimonious way by investigating the effects of the composite predictors (i.e., principal components) on the criterion variables at different levels, as well as the effects of the predictor variables on the criterion variables via the composite predictors. Some possible extensions of the proposed method are also suggested.

55 Psychoperiscope

Joshua Chiroma Gandi, *University of Jos, Nigeria*

Coping refers to effort towards mastering demands pose by harm, threat or challenge appraised/perceived as taxing available resources. It could be in terms of problem-focused versus emotion-focused as well as behavioral (what you do) versus cognitive (what you know) dimensions. Cognitive coping strategies regulate emotions through cognitions that are inextricably associated with human life. Therefore, assessing its effect in the relationship between particular life events and perceived quality of life competes for attention. Hence based on Gandi Psychometric Model, using a coined nomenclature which combines psychometrics and periscope, "psychoperiscope" was conceptualized and developed. Predictive correlational (cross-sectional) design, which in this case involves systematic investigation of the relationship between particular life events and quality of life, was adopted in the study. Focusing on the effects of cognitive coping strategies, this correlational design helps to predict the variance of perceived quality of life based on the variance of particular life events. The scale was pilot-tested on 30 target participants selected by stratified random sampling alongside their respective family member significant others (n=30) and clinical practitioner significant others (n=30). Item validation resulted to a 3-version scale with respective 28 items showing that psychoperiscope is a psychometrically suitable scale for assessing covariance between particular life events and quality of life. It is a research instrument which also serves as a screening and diagnostic tool. Following the meaningful and useful data it elicited in this study, psychoperiscope would effectively generate more optimal and robust data if complemented with an experimental study using appropriate equipment.

56 WEB GESCA: Web-based software for generalized structured component analysis

Seungman Kim, *Texas Tech University, United States*
Kwanghee Jung, *Texas Tech University*
Heungsun Hwang, *McGill University*

Generalized structured component analysis (GSCA), which is a component-based approach to structural equation modeling, has been extensively enhanced in terms of data-analytic capability and flexibility as well as computational efficiency. To facilitate the diffusion and wide adoption of this approach by SEM users, we have developed a free, online software program, named WEB GSCA (www.sem-gesca.org/webgesca), using R Shiny and Shiny Dashboard. This software provides a web-based graphical user interface that allows users to easily build a model by checking checkboxes in model specification tables, to run GSCA, and to view results on the program window. The software also has several advantages. First, users can execute the software on the web without the need of downloading it to their computer. Second, they can execute the software on any operating system as long as they have access to the Internet. Third, the web-based software is continually updated in real time, thus users can always use its latest version. We will illustrate how to use the software.

57 Hybrid-GIMME combining uSEM and SVAR

Lan Luo, *University of North Carolina at Chapel Hill, United States*
Cara Arizmendi, *University of North Carolina at Chapel Hill*
Kathleen Gates, *University of North Carolina at Chapel Hill*

Researchers with time series data often have to decide how to model contemporaneous, or lag-0, relations among variables. The current Group Iterative Multiple Model Estimation (GIMME) algorithm allows for the search of unified Structural Equation Models (uSEM), which includes directed contemporaneous relations. However, because the paths in uSEM are directional, it can introduce model identification issues if bidirectional contemporaneous relations exist. Another possible method, Structural Vector Autoregression (SVAR) allows for bidirectional contemporaneous correlations among errors in a VAR model. Molenaar (2018) introduced the hybrid-VAR, an approach that allows for both directed and correlational contemporaneous relations. This poster introduces an extension of this work, Hybrid-GIMME. Hybrid-GIMME is a data-driven approach that provides a more flexible model. We focus on the technical differences between the two methods and end with both an empirical and a simulated example demonstrating finding the potential relations between time series variables using Hybrid-GIMME.

58 Burnout syndrome in Brazilian university professors and academic staff members.

Fernanda Ludmilla Rocha, *University of São Paulo, Brazil*
João Marôco, *Institute of Psychological, Social and Life Sciences (ISPA)*
Maria Helena Palucci Marziale, *University of São Paulo*

Juliana Alvares Duarte Bonini Campos, *São Paulo State University*

The aims of this study were to evaluate the psychometric properties of Copenhagen Burnout Inventory Brazilian version (CBI-Br) in a sample of university professors and academic staff members. A cross-sectional study design was used. The sample was composed by 676 workers. The psychometric properties of the CBI-Br were analyzed using Confirmatory Factor Analysis (CFA). The reliability of the items was estimated using Cronbach's α and Composite Reliability (CR). The overall weighted scores of the instrument were compared to the sample's demographic characteristics using ANOVA (significance level=5%). About the sample, 56.2% were women; 54.7% were professors; 48.05 years were the sample's age average; 380 worked for up to 15 years in the universities and 94.6% worked more than 40 hours per week. The CFA showed an acceptable overall fit of a three-factor model with 18 items. The convergent validity, CR and the Cronbach's α for all CBI-Br factors were adequate and it was observed the strict measure invariance of the refined model. Gender (women) was a social determinant for the occurrence of BS unlike the function. The correlation between the worked hours per week and the working time and Burnout was negative and statistically significant ($p<0.05$) only to women. The results of this study provided evidences of the validity and reliability of CBI-Br to the measurement of BS in Brazilian university professors and academic staff members. Besides, the CBI-Br may represent an important tool for the diagnosis of psychosocial risks related to the BS in the academic environment.

59 Psychometric analysis of the Work Limitation Questionnaire in university workers

Fernanda Ludmilla Rocha, *University of São Paulo, Brazil*
Samuel Andrade de Oliveira, *University of São Paulo*
João Marôco, *ISPA*
Juliana Alvares Duarte Bonini Campos, *São Paulo State University*

To evaluate the psychometric properties of the Work Limitation Questionnaire (WLQ) applied to a sample of university professors and academic staff members of Brazilian universities. A cross-sectional study design is used (n=393). The psychometric properties of the WLQ were analyzed by estimating the psychometric sensitivity, the factorial validity, the convergent and discriminant validity, and the reliability of the model. The factorial validity was performed using Confirmatory Factor Analysis with Maximum Likelihood estimation method. The reliability was estimated using standardized Cronbach's α and Composite Reliability. The overall score of the instrument was calculated using the matrix of factor score weights and was compared to the sample demographic characteristics using ANOVA (significance level=5%). The sample's demographic data showed an average age of 47.88 years (SD=9.217); 62.8% were women; 55% were academic staff members; 66.2% worked for up to 10 years and 95.2% worked more than 40 hours/week. The CFA showed a refined model with adequate overall fit (x2/df=2.785; CFI=.956; TLI=.950; RMSEA=.067), as well as the hierarchical second-order model (x2/df=2.776; CFI=.956; TLI=.950; RMSEA=.067), with strong contribution of the factors TM, MI and OD ($\beta >.90$) to the general concept of presenteeism. It was observed adequate convergent validity and reliability of the instrument for the sample. The comparisons of the overall weighted scores of WLQ factors between gender, function, worked hours and working time were not statistically significant. In conclusion, the WLQ is a valid and reliable instrument to the assessment of presenteeism in Brazilian university professors and academic staff members.

60 Sample size planning for standardized SEM parameters given assumption violations

Ana Simoes, *University of California, Merced, United States*

A method for planning sample size is developed with the goal of achieving sufficiently narrow confidence intervals of standardized Structural Equation Modeling (SEM) parameters when model misfit and nonnormality of data are present. First, sample size planning methods tend to focus on the overall model fit through null hypothesis significance tests and fit indices. However, model fit does not implicate in relevant effects among variables in the model. Therefore, the method focuses on narrow confidence intervals of parameters to ensure the precision of such effects. Next, since model correctness and normal distribution of data are assumptions that are often violated in the social sciences, the method accounts for these limitations by using robust maximum likelihood estimation in its procedure. Finally, methods for planning sample size are usually developed using unstandardized parameters, which can be challenging to interpret and to specify. Therefore, the proposed method emphasizes sample size planning in the standardized context. Using the method, sample size was calculated for different conditions varying according to model complexity (i.e. types

of model included one confirmatory factor analysis and two SEM models), the parameter of interest, the desired confidence interval width, the degree of misfit and the degree of nonnormality of data. Simulation studies were conducted to assess the effectiveness of the method.

61 Access to psychological services and its influence on learners' career choice.

Jesse Ashley, *University of Johannesburg, South Africa*

In the year 2017 and 2018, a total of about 750 learners seek guidance and counselling support (applying psychometric test batteries) to aid them with their subject choice and future career options. Various research have shown that guidance and counselling services are very important tools in human development especially during adolescent stage. A lack of guidance and counselling among adolescents may result in the increase of unpleasant outcomes in the society. According to the South Africa's National Policy for Integrated Career Development System, every citizen (learners) must have access to career guidance. We ask if there is significant relationship between accessing guidance and counselling services on learners study life and attitude. Also, we ask if there is a significant relationship between accessing guidance and counselling services and career choices. Thirdly, if there is significant relationship between learner's attitude towards studies and career choice. The target population in this research was identified as Grade 10 and 12 learners who have received various guidance and counselling services from Sci-Bono. Consent forms were given to learners below the age of 18 to be given to their parent as learners above age 18 were also given consent forms to fill. The study shows that accessing guidance and counselling services has an impact on learners' school life. It is imperative therefore, that schools and the department of education see the importance of extending the services of guidance and counselling services to schools.

62 Adaptation and validation of the PERMA profile

Yuri Chávez Luque, *Universidad Nacional de San Agustín de Arequipa, Peru*

The research sought the analysis and adaptation of the profile PERMA test from Kern, Waters, Adler & White; of the year 2014, in a Peruvian and Colombian context, this instrument is framed in positive psychology and the five major factors mentioned by Seligman in this theory, these are: Positive emotion (P), Commitment (E), Relationships, (R) , Meaning (M), Achievement (A); and study both the psychometric properties of the instrument, and the theory is very important. The test obtained a high level in its validity of content being reviewed by experts who reviewed the items agreeing with all those raised, in terms of their criterion validity, because in the absence of a test that measures the same was compared with the scales of psychological well-being of José Sánchez-Cánovas (EBP), comparing it only with the scale that concerns only the factor of psychological well-being, obtaining a correlation of 0.7 with the total of the PERMA scale. It was also compared with the Diener Subjective Wellbeing Scale (1994), obtaining a correlation of 0.5. Regarding the construct validity, the exploratory and confirmatory factor analysis was applied, finding only three of the five factors proposed by the theory; Finally, the Item Response Theory was applied with the Rasch model.

63 Validation of a Satisfaction scale for ambulatory medical consultations

Edward Mezones-Holguín, *Universidad San Ignacio de Loyola, Peru*
Miguel G. Moscoso, *Universidad Peruana Cayetano Heredia*
David Villarreal-Zegarra, *Universidad Peruana Cayetano Heredia*
Ronald Castillo, *Universidad del Pacifico*
Luciana Bellido-Boza, *Intendencia de Investigación y Desarrollo, Superintendencia Nacional de Salud*

Background: Patient satisfaction is a vital indicator for policy-makers to improve or change the offer of health services. In Peru, the complexity of the health system calls for a unified measure of satisfaction that allows to perform fair assessments and comparisons. Objective. To design and assess the psychometric properties of a scale that measures satisfaction with ambulatory medical consultations in health centers from the Peruvian health system. Methods. Instrumental study using a dataset from the National Survey of Satisfaction of Health Users, where an initial scale of 19 items constructed by bibliographic review, expert judgment, and pilot study was applied. We performed an exploratory factor analysis (EFA) and a confirmatory factor analysis (CFA) using robust maximum likelihood estimations to assess the internal structure of the scale. Finally, we performed measurement invariance analysis and assessed reliability with McDonald's omega coefficient (ω). Results. We analyzed 13814 observations, randomly divided in two subsamples, for the EFA which resulted in a three-factor model with 18 items that showed acceptable goodness-of-fit indexes (CFI = 0.945, TLI = 0.937, SRMR = 0.036) in the CFA. These three factors were satisfaction with administrative processes,

infrastructure, and medical attention. We found strong invariance for age, sex, educational level and area of residence, and partial invariance for institution type. All reliability coefficients were deemed as good ($0.86<\omega<0.92$). Conclusion. This scale designed to measure satisfaction with the ambulatory medical consultation presents evidence of validity and reliability, and measurement invariance at different levels of service in a national-based sample.

64 Certainty-based marking on multiple-choice items: A decision-making perspective

Qian Wu, *University of Leuven, Belgium*
Rianne Janssen, *KU Leuven - University of Leuven*

When responses to multiple-choice items consist of selections of single-best answers, it is not possible for examiners to differentiate between responses that are a product of knowledge and those that are largely a product of uncertainty. Certainty-Based Marking (CBM) is one testing format to obtain such information, which additionally requires examinees to rate their degree of certainty on the selected single-best answers. Responses are scored on the correctness of the chosen option and the degree of certainty examinees have in their choice. The expected score is maximized if examinees truthfully report their level of certainty. However, prospect theory (Kahneman & Tversky, 1979) states that people do not always make rational choices of the optimal outcome due to varying risk preferences. The present study looks into response behaviors of 334 first-year students of kinesitherapy on six exams with CBM, and examines whether there is a bias regarding the discrepancies between the actual accuracy rates and the subjective certainty ratings at the item and test levels as well as individual differences therein. Results from the score group analysis show that overall certainty ratings were positively associated with accuracy rates. However, lower ability students tended to overestimate their certainty levels, whereas higher ability ones underestimate, even on tests with lower difficulty items. This shows that students' responses were affected by their risk preferences and scoring rules. Moreover, female students on average had higher accuracy rates and certainty ratings, but no significant gender difference was found regarding the mis-calibration of certainty levels.

65 Psychometric properties of the SSS in an Aboriginal population

Pedro Ribeiro Santiago, *The University of Adelaide , Australia*
Rachel Roberts, *The University of Adelaide*
Lisa Smithers , *The University of Adelaide*
Lisa Jamieson , *The University of Adelaide*

Due to a history of colonization, Aboriginal and Torres Strait Islander (ABTSI) became one of the most disadvantaged groups in Australia. The decades of assimilation policies, which culminated in the removal of Aboriginal children from their parents (i.e. the "Stolen Generations"), disassembled their society and mitigated the social support derived from their communities. There are no psychological instruments validated to measure social support in Aboriginal Australians. The aim of the current study was to evaluate the validity and reliability of the Social Support Scale (SSS) in a population of Aboriginal women. The SSS is a 4-item instrument developed to measure the emotional, appraisal, informational and instrumental dimensions of social support. Data was collected from the Baby Teeth Talk Study (n=367), an RCT conducted among Aboriginal women in South Australia. The psychometric properties were evaluated with Graphical Loglinear Rasch Models, which extend the Rasch Model to incorporate uniform local dependence (LD) and differential item functioning (DIF). Overall fit to a GLLRM (chi2 (25) = 22.2, p=0.625) was found after the inclusion of LD between items 3 and 4 (gamma-avg=0.66). Unidimensionality was confirmed (gamma-obs=0.75, gamma-exp=0.77, p=0.163) and items had no DIF. The SSS displayed good reliability (R sample=0.84), probability of person separation (P sample=0.78), while targeting was poor (TTI sample=0.49). The current study consisted of the first validation of a psychological instrument to measure social support in ABTSI. The results show that the SSS is a valid and reliable psychological instrument that can be applied among Aboriginal women.

Author Index