IMPS 2015

International Meeting of the Psychometric Society

Beijing Normal University, China

Abstract Book: Posters

See Also:

- Abstract Book: Talks
- Program Book (printout included in conference package)
- Schedules for Monday, Tuesday, Wednesday (printouts included in conference package)
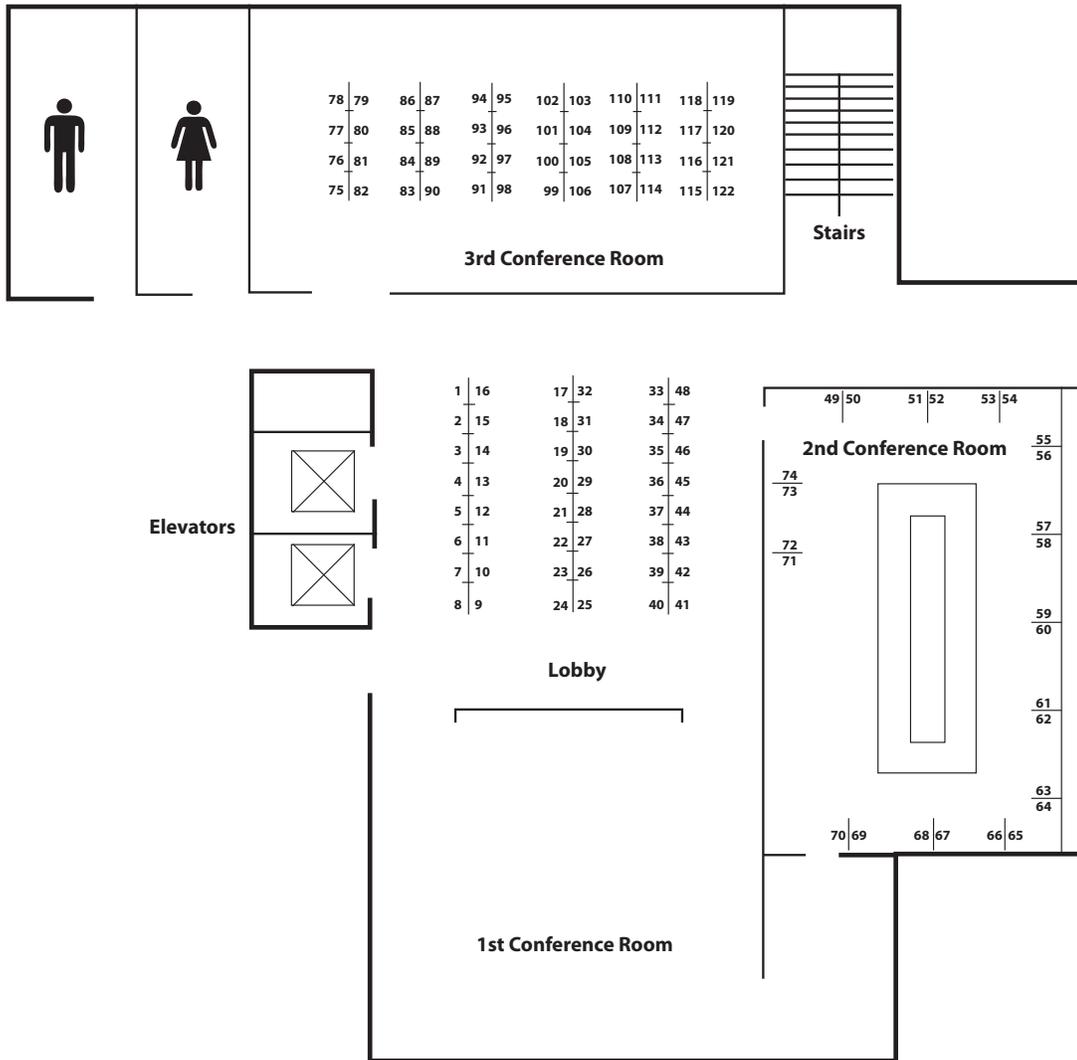- Conference Website

Updated on July 3, 2015

Welcome Reception · Poster Session

Monday, July 13, 2015
4:30 p.m.-6:30 p.m. (16:30-18:30)

Location: Jingshi Xuetang - B1 Floor (X-B1)

78 79	86 87	94 95	102 103	110 111	118 119
77 80	85 88	93 96	101 104	109 112	117 120
76 81	84 89	92 97	100 105	108 113	116 121
75 82	83 90	91 98	99 106	107 114	115 122

3rd Conference Room

Stairs

Elevators

1 16	17 32	33 48	49 50 51 52 53 54
2 15	18 31	34 47	
3 14	19 30	35 46	55/56
4 13	20 29	36 45	74/73
5 12	21 28	37 44	57/58
6 11	22 27	38 43	72/71
7 10	23 26	39 42	59/60
8 9	24 25	40 41	61/62

2nd Conference Room

63/64

70 69   68 67   66 65

Lobby

1st Conference Room

Poster Presenters

1.	Sakaya Arai	41.	Keke Lai	81.	Shaojing Sun
2.	Lü Bo	42.	Mina Lee	82.	Xiao-Jian Sun
3.	Ralph Carlson	43.	Sora Lee	83.	Teck Kiang Tan
4.	Hyun Jung Chae	44.	Young-Sun Lee	84.	Yun Tang
5.	Fu Chen	45.	Cheng-Hsuan Li	85.	Jian Tao
6.	Shu-Ying Chen	46.	Hui Li	86.	Bing Tong
7.	Young Il Cho	47.	Hua-Xia Li	87.	Joyce Chang
8.	Jinnie Choi	48.	Xiaoyue Li	88.	Peng Wang
9.	Chun Yu	49.	Yujun Li	89.	Wen-Yi Wang
10.	James P. Clifton	50.	Miaoyun Li	90.	Yinan Wang
11.	Zhongmin Cui	51.	Kai-Li Liang	91.	Kimberly Ward
12.	Stephan Daus	52.	Lu Liang	92.	Weidong Wang
13.	Amanda J. Fairchild	53.	Chuan-Ju Lin	93.	Yi-Jhen Wu
14.	Ania Filus	54.	Jie Liu	94.	Shibei Xiang
15.	Luise Fischer	55.	Xian-Wei Liu	95.	Xin Xin
16.	Lingyun Gao	56.	Xiao-Xiao Liu	96.	Kuan Xing
17.	Zolio Emilio García Batista	57.	Ying Liu	97.	Xinli Zhu
18.	Kiero Guerra Peña	58.	Yanan Liu	98.	Qinfang Xu
19.	Fei Guo	59.	Yanlou Liu	99.	Michio Yamamoto
20.	Lei Guo	60.	Yuan Liu	100.	Chenchen Yang
21.	Qingke Guo	61.	Zhenqiu (Laura) Lu	101.	Heewon Yang
22.	Yage Guo	62.	Fen Luo	102.	Ji Seung Yang
23.	Jilian Han	63.	Xin Luo	103.	Xi Yang
24.	Woori Han	64.	Jingyao Mai	104.	Yuzhu (June) Yang
25.	Ye He	65.	Sunhong Min	105.	Qin Yi
26.	Xiaoting Huang	66.	Hong-Hua Mu	106.	Yi Tian
27.	Gao Jian	67.	Tomoya Okubo	107.	Tian Yu
28.	Yanxia Jiang	68.	Xiangzi Ouyang	108.	Jianlin Yuan
29.	Xiong Jianhua	69.	Fan Pan	109.	Ma Yue
30.	Jiaxuan Du	70.	Yiqin Pan	110.	Yuheng Huang
31.	Zhang Jinhua	71.	Theresa Rohm	111.	Duan Zhang
32.	Kentaro Kato	72.	Yuan Rongman	112.	Qun Zhang
33.	Keonseob Kim	73.	Hye-Eun Seok	113.	Xue Zhang
34.	Kyoungshin Kim	74.	Xuejun Shen	114.	Mingying Zheng
35.	Kyung Yong Kim	75.	Yun Shi	115.	Shan Zhiyan
36.	Meereem Kim	76.	Shu-Chuan Shih	116.	Hao Zhou
37.	Nana Kim	77.	Mohammad Shoraka	117.	Qun Zhou
38.	Sewon Kim	78.	Ding Shuliang	118.	Wanjing J. Zhou
39.	Renske E. Kuijpers	79.	Xiaojuan Song	119.	Yuan Zhou
40.	Gökhan Kumlu	80.	Mei Su		

Contents

1 A conventional method for analysis of characteristics of test items

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

Academic examination data are usually analyzed using Item Response Theory (IRT). IRT, which is also called latent trait theory, is a method for estimating latent traits (e.g., academic proficiency). In IRT, characteristics of test items are described with a few parameters, such as discrimination parameters and difficulty parameters. IRT is widely used in many educational testing applications, however, the computational procedures are not so simple and it requires a large number of examinees to obtain reliable estimates of the parameters. In this study, we propose a simple but effective method for analyzing characteristics of test items. The method utilizes the idea of multiple correspondence analysis and signal detection theory. Our method provides indices of item difficulty and item discrimination. It contributes to the process of item analysis, so it helps to evaluate the effectiveness of an item and improve tests. It also yields diagnostic information on students. In our presentation, we analyze a real data set of practical reading test and compare the results of IRT with those of our new method. Also, we discuss the relationship between IRT and our method. [29942]

2 Adolescents' academic achievement and emotion: The moderating role of parental attitude and behavior

Lü Bo, *Beijing Normal University, China*
Luo Liang, *Beijing Normal University, China*

Drawing on the background of positive psychology, more and more studies have focused on chidren's wellbeing in the context of education. Studies examining the relationship between children's academic achievement and wellbeing have shown conflicting results. So there may be some moderators in this relationship. Early studies found that parenting has an important effect on children's wellbeing and academic achievement, but most studies just focused on the main effect. The present study investigates the role of parents' attitude and behavior towards schools as a potential moderator of the relationship between children' academic achievement and emotional wellbeing (both positive and negative emotions). 419 students, who were in 4th to 6th grade, and their parents from two elementary schools in Liaocheng of Shandong province, took part in our study. Students' academic achievement was operationalized by Grade Point Average (GPA) in three subjects. Results indicated that both the attitude and behavior of parents functioned as moderators of the relationship between academic achievement and students' positive emotion. The association between academic achievement and positive emotion was only found in the group of students whose parents have a positive attitude and behavior towards school. The behavior of parents also moderates the relationship between academic achievement and students' negative emotion in a similar way. This study can explain the conflicting results in previous research. The attitude and behavior of parents may play an important role in the relationship between academic achievement and students' wellbeing. [29677]

3 Sociocultural-linguistic diversity and performance on nested multilevel measures

Ralph Carlson, *The University of Texas Pan American, USA*
Hilda Medrano, *The University of Texas Pan American, USA*
Julissa Barboza, *The University of Texas Pan American, USA*
Yoonkyung Chung, *The University of Texas Pan American, USA*
Carlo Rafael Flores, *McAllen Independent School District, USA*

Variance in nested multilevel measures is transitive, that is, it flows from bottom of a nested multilevel/hierarchical structure to the apex. Thus, analysis and interpretation of nested multilevel/hierarchical measures must be from the bottom up. When there are significant discrepancies/differences within a lower level/stage/measure of a nested multilevel/hierarchical structure, all successive levels are not interpretable due to lack of "cohesion" or ambiguity; therefore, there is a problem with loss of information. The current study presents a model that adjusts for nested effects in nested multilevel/hierarchical measures and, thus, provides a solution by alleviating the loss of information due to a lack of "cohesion" across successive levels of a nested multilevel/hierarchical structure. This model was tested with a data set obtained on the Wechsler Intelligence Scale for Children from a random sample of 9 year-old bilingual Hispanic children. [29539]

4 Generalizability approaches to evaluating measurement error sources and reliability of portfolio assessments in Korean elementary schools

Hyun Jung Chae, *Yonsei University, South Korea*

Su Hyang Hong, *Gyeongin National University of Education, South Korea*
Guemin Lee, *Yonsei University, South Korea*
Seungeun Chung, *Yonsei University, South Korea*

The purpose of this study was to explore error sources and the reliability of the portfolio assessments in Korean elementary schools. The portfolio assessments used in this study were from the "Wise life" course, that is a required subject area for 2nd and 3rd grade elementary school students. This study investigated the estimation of reliability and the optimal measurement conditions with various tasks and raters to improve reliability in portfolio assessments based on generalizability theory. Data were gathered through portfolio assessments of second grade elementary school students. Three evaluation tasks were given to students, and two raters (elementary school teachers) scored the outcomes of all students. In conclusion, it was more efficient to increase the number of tasks rather than that of raters in order to satisfy the optimal conditions for the portfolio assessments. [29574]

5 Applying the rule space model to develop a learning progression of thermochemistry

Fu Chen, *Beijing Normal University, China*
Tao Xin, *Beijing Normal University, China*
Yanfang Guo, *Beijing Normal University, China*

We used the rule space model, a kind of cognitive diagnostic model, to measure learning progression of thermochemistry in senior high school students. We extracted five attributes and specified their hierarchical relationships to model the construct of thermochemistry in four levels by the hypothesized learning progression. For this study, we developed 24 items addressing the attributes of exothermic and endothermic reactions, chemical bond and heat quantity change, reaction heat and enthalpy, thermochemical equation and Gass' Law to a sample base of 694 senior high school students taught in 3 schools across 2 cities. Results based on the rule space model analysis indicated that: (1) the items of the test developed by the rule space model were of high psychometric quality in terms of difficulties, discriminations, reliabilities and validities; (2) the rule space model analysis classified students into 7 different attribute mastery patterns; (3) the sequence of each attribute's average mastery probability from low to high corresponded to the hierarchy model for the attributes; (4) the initial hypothesized learning progression was modified by the attributes mastery patterns and the learning paths in a more precise and detailed way. [29709]

6 Detection of item compromise in computerized adaptive testing

Shu-Ying Chen, *National Chung Cheng University, Taiwan*
Pin-Wun Chen, *National Chung Cheng University, Taiwan*

For high stakes testing programs, no matter how tightly item exposure is controlled, it is still possible for some examinees to obtain test information from previous test takers and have their test scores seriously inflated. The examinees' scores would not be seriously inflated if the compromised items could be effectively detected and excluded from test administration. The proportion correct could be a good index for detecting compromised items. In CATs, the probability that an examinee can answer a selected item correctly is close to 0.5 because the difficulty of the selected item is close to the examinee's ability. Thus, the proportion correct for each item should be about 0.5 no matter what item difficulty is. When item sharing occurs, however, the proportion correct for a compromised item could be far above 0.5. Thus, based on changes in the proportion correct, compromised items could be identified. A sequential procedure was proposed by Zhang (2014) to detect compromised items by examining the changes in proportion correct. This procedure, however, has limited power for detecting compromised items when item sharing occurs at the early stages. Since the earlier the item is compromised, the more severe the damage would be, compromised items should be effectively identified, especially for those compromised at the early stages. To have compromised items effectively detected at the early stages, this study attempts to take response times into account in cheating detection. The effects of the proposed procedure on detecting compromised items is thoroughly investigated. [29661]

7 Evaluating individual model fit in mixed effects models: Identifying individuals showing different learning processes from those of the predicted growth model

Young Il Cho, *Sungshin Women's University, South Korea*
Yujung Jo, *Sungshin Women's University, South Korea*
Jihyun Kim, *Sungshin Women's University, South Korea*

Mixed Effects Models (MEMs) have used overall model fit indices such as the Akaike Information Criterion (AIC; Akaike, 1987) and the Bayesian Information Criterion (BIC; Schwarz, 1978) to evaluate the appropriateness of a model fitted to longitudinal data. In contrast to overall model fit, an individual fit statistic may detect individuals whose responses are not properly fitted by the predicted

model. In this study, a within-individual deviation score (within score) was computed and used to evaluate model fit to empirical repeated measures on learning process in quantitative tasks at an individual level. Based on a simulation study, Cho & Blozis (in preparation) demonstrated that the within score had an expected mean and standard deviation of a chi-square distribution with corresponding degrees of freedom. The within score showed that the learning process of most participants in the sample followed a negative accelerated exponential function but some individuals followed a different trajectory. In conclusion, even though overall fit indices showed that the model fits the data well, the individual fit indices indicated that the model was not suited for several individuals. The within score accurately evaluated model misspecification at an individual level. Additionally, limitations of alternative individual fit indices were discussed. [29955]

8 Some psychometric heuristics for evaluating performance of assessment items

Jinnie Choi, *Rutgers University, USA*
Ravit Golan Duncan, *Rutgers University, USA*

In educational and psychological assessment, the measurement items are important tools to elicit evidence of where the examinees are on the scale of the construct in focus. Evaluating the performance of these items is a crucial step towards a more reliable and valid assessment. Pilot testing can provide rich data sets that can be used to evaluate item functioning. However, there is a variety of psychometric information on item quality. Coordinating and weighting this information in order to determine item overall quality is not trivial. Another challenge is deciding how to respond, in terms of item revision, to the available psychometric information. A statistical measure may highlight a problematic aspect of the item, but it does not dictate a specific revision in terms of language or structure of the problematic item. Often a series of consecutive pilots are needed to optimize item functioning and progress is not always in a forward direction; at time an evidence-driven revision could worsen the item's functioning on some dimensions. In this paper, we describe the psychometric heuristics we use for developing and improving assessment items. We illustrate these heuristics through examples from an iterative research project on a multi-dimensional science learning progression. We attempt to provide useful insights regarding three interrelated questions: (a) what evaluative indices of item quality and performance can be used? (b) how do we assemble and interpret a variety of information from item evaluation? (c) what are the affordances and constraints of various solutions to common item problems? [29918]

9 Controlling MAX test overlap in computerized adaptive testing

Chun Yu, *Taiwan*
Shu-Ying Chen, *Taiwan*

To date, Computerized Adaptive Testing (CAT) has been used widely all over the world. How to maintain test security is an important issue. Test overlap is an important index of test security in CAT. To control test overlap in CAT, Chen (2012) developed the MGT procedure. However, the MGT procedure can only control "general test overlap", i.e., "average" test overlap for between current and previous examinees. In MGT, although average test overlap can be controlled at an appropriate level, MAX test overlap cannot be well controlled, sometimes the MAX test overlap rate can be as high as 1.0. That is, for some examinees, their test items are exactly the same as those taken by previous examinees. To deal with this shortcoming of the MGT procedure, Chen (2013) developed the MGT_M procedure to control not only the general test overlap but also the MAX test overlap. Nevertheless, the MGT_M procedure did not perform well for controlling MAX test overlap, especially for stringent test conditions. This study attempts to modify the MGT_M procedure to improve its ability to control the MAX test overlap. [29577]

10 Skewed within-class mixture distributions in latent growth mixture modeling: An assessment of specification errors and class enumeration

James P. Clifton, *University of California, Merced, USA*
Sarah Depaoli, *University of California, Merced, USA*
Keke Lai, *University of California, Merced, USA*

Recent advances have made it possible to model mixture components within Growth Curve Models (GCMs) and Growth Mixture Models (GMMs) using robust, skewed mixture distributions. This feature allows for flexibility in handling non-normality in longitudinal data, through manifest or latent variables, by modeling skewed or heavy-tailed latent classes rather than assuming normality. In this simulation study, we assessed the potential under- and over-extraction of latent classes in a growth model when the underlying data follow normal, skewed-normal, or skewed-t distributions. Specifically, we assessed model comparison, fit, and classification of correctly specified and misspecified models through the AIC, CAIC, BIC, aBIC, LMR-LRT, BLRT, and entropy. We manipulated

several design factors, including: (a) type of model (2-levels: GCM versus GMM), (b) mixture model distribution (3-levels: normal, skewed-normal, or skewed-t), (c) latent class separation (2-levels: low versus high), and (d) relative latent class size (2-levels: equal versus unbalanced). All data were generated using Mplus version 7.2 with 500 replications per cell and a sample size of n = 1,000. Results revealed that each of the information criteria and LRT-based measures accurately recovered the correct class structure when classes were well separated, but tended towards under-extraction when class separation was low. In addition, each of the model comparison and fit measures tended towards over-extraction when the data consisted of a single class characterized by a skewed-t or skewed-normal distribution, and the model was mis-specified as consisting of normal mixtures. Implications of findings and future research directions are discussed in the presentation. [29617]

11 Practical considerations in choosing an anchor form for equating

Zhongmin Cui, *ACT, USA*
Yong He, *ACT, USA*

In the random groups design, anchor forms play an important role in equating new test forms to the base form and in maintaining the score scale. When there are several candidates for the anchor form, test developers need to make a decision on which one to choose. The research literature shows plenty research on anchor items but not much on anchor forms, although the role of anchor forms on maintaining score scales is no less important than anchor items. In this study, we attempt to explore factors that may help in anchor form selection. These factors include, but are not limited to, reliability of test forms, scale score distributions, standard error of equating, gaps and runs in the raw to scale score conversion tables, and similarity of raw score distributions. We conduct a simulation based on item parameter values from real data. Multi-year administration is simulated and the equipercentile method is used for equating. The aforementioned factors are studied one by one as well as in pairs to see which factor or pairs of factors can help to choose anchor forms that result in most accurate equating relationships. The accuracy of equating is evaluated by computing the root mean square difference between the simulated operational equating relationship and the true equating relationship. Putting all results together, we attempt to make some recommendations to practitioners on constructing and choosing anchor forms. Test score users would also benefit from the results of this study by potentially getting more accurate test scores. [29692]

12 Explanatory item response perspectives on large-scale assessments

Stephan Daus, *University of Oslo, Norway*

In the context of further validation of large-scale assessments, the current study uses an explanatory item response approach to hierarchically decompose the item difficulty of the TIMSS (Trends in International Mathematics and Science Study) science component into content domain structure and curricular information. The TIMSS science item pool consists of item families structured around common domains (e.g., biology, physics) which are further divided into topics and objectives. Within the large framework of non-linear mixed models, both persons and items errors are assumed to be random, with the implied item hierarchy modelled using either fixed effects or a further nesting of random effects. In addition to the item responses and content categories, information on the curricular content coverage is included in TIMSS. This information can be incorporated into the item response model. Curricular information at the national level (intended curriculum) represents an item predictor, while curricular information at the classroom level (implemented curriculum) varies across persons and thus represents an item-person interaction predictor. The findings have implications for construct validity with respect to usage of results from large-scale assessments; in particular as standards for educational testing require Opportunity to Learn. This means that curricular coverage ought to be considered when interpreting such high-stakes assessments. [29907]

13 Evaluating the efficacy of estimation strategies for discrete-time survival mediation effects: A simulation study comparison of the product of coefficients and a potential outcomes framework

Amanda J. Fairchild, *University of South Carolina, USA*
Chao Cai, *University of South Carolina, USA*

Though the mediation model is causal in theory, non-randomized studies often preclude inferring causation. Recent methodological advances have made progress in improving causal interpretation of model parameters however. These potential outcome approaches are an important new tool for the field. This proposal extends exploration of model estimation for Discrete-Time Survival Mediation (DTSM) effects by comparing Imai et al.'s (2010) causal inference approach for estimating mediation to the original specification proposed in earlier work. We conducted a simulation study to assess differences in estimation strategies across a $2 \times 3 \times 4 \times 4 \times 2 \times 2$ factorial design examining a variety of effect size, sample

size and time-point combinations to compare power, Type 1 error, CI coverage, MSE and accuracy of parameter estimates across approaches. Utilizing the R 'mediation' package, we invoked the 'glmer' function with a logit link to model the survival data structure via a random effect component to obtain estimates of Average Causal Mediated Effects (ACMEs) in each parameter combination. Under the assumption of sequential ignorability and linearity (which holds in logit form), it has been proved that ACME is identified by the usual product of coefficients. Initial study results indicate that given small variance estimates for the random effect, the 'mediation' package yields ACMEs comparable to estimates garnered from the original simulation study. However, higher variance estimates in the random effect component introduce inaccuracy into individual parameter estimates, thus compromising the ACME. Implications are discussed. [29649]

14 Multi-group latent class analysis as an alternative approach to analyze parental preferences for service delivery across countries

Ania Filus, *The University of Queensland, Australia*
Alina Morawska, *The University of Queensland, Australia*

Multi-Group Latent Class Analysis (MG-LCA) is a valuable procedure for cross-cultural studies evaluating parental preference for service delivery, yet it is rarely used in implementation research. To illustrate the utility of this approach in implementation research we applied it to the data from the International Parenting Survey designed, among others, to investigate parental preferences for formats of receiving parenting information in Australia and China. This is a novel application of MG-LCA in the field of implementation science. 1172 parents (583 from Australia; 589 from China) participated in the study. Measures were available in English. The translation back-translation method was used to prepare the Mandarin version of the survey (Brislin, 1972). The MG-LCA revealed 2 groups of parents-consumers in each country: those who prefer (C1) face-to-face and (C2) light-touch parenting interventions. The Wald tests indicated significant differences in item endorsements in each class between the countries (lack of measurement invariance). For C1, Chinese parents were more likely to endorse face-to-face modalities as compared to Australian parents. For C2, Australian parents were more likely to endorse light-touch interventions as compared to Chinese parents. Since item-response probabilities were not equal across the groups, following the Collins & Lanza (2010) recommendation we did not test for similarities/differences in

latent class prevalences across groups. Further analyses indicated that in Australia parents who consider parenting as a private matter were less likely to be in a group preferring face-to-face interventions (OR = .91, 95% CI (.85-.97)). No significant predictors of class membership were found in China. [29601]

15 Looking for the true link: Comparing different linking methods and their related linking errors on Rasch scaled competence data from the NEPS/grades 5 and 7

Luise Fischer, *Leibniz Institute for Educational Trajectories, Germany*
Claus H. Carstensen, *Leibniz Institute for Educational Trajectories, Germany*
Theresa Rohm, *Leibniz Institute for Educational Trajectories, Germany*

Linking data is an important topic in investigating educational trajectories. As such, our aim is to appropriately link the growing number of competence data from the German National Educational Panel Study (NEPS). In the present study, different approaches on linking two successive tests measuring mathematical competence are compared; aiming to find a valid prototypical procedure for future link studies to come. With our data being scaled using Masters' Partial Credit Model (PCM; 1982), restrictions on the data are very tight, but allow for the assumption of specific objectivity. Consequently, if the PCM holds for the data, the comparison and linking of participants' performance across different measurement points is valid. In a longitudinal design, participants were administered a test on mathematical competence in grade 5 (g5; n=6112, items=24) and grade 7 (g7; n=6194, items=23), each containing 6 link items. Due to linking purposes, an extra sample in grade 7 was administered a merged version of the g5-test and the g7-test (g5g7; n=581, items=47). Since the administered tests (g5, g7, g5g7) share a common pool of items, different linking methods are compared: First, g5-test and g7-test are linked directly via fixed parameters and mean/sigma-linking. Second, as an indirect way of linking, g5g7-test is scaled and item estimates are transferred to the g7-test afterwards. And third, all tests (g5, g7, g5g7) are scaled simultaneously through concurrent calibration. Of particular interest hereby is the linking error, it is a way of computation and it is magnitude within the different linking procedures. [29641]

16 An innovative CAT algorithm for adaptive test batteries

Lingyun Gao, *ACT, USA*
Lisa Gawlick, *ACT, USA*
Nancy Petersen, *ACT, USA*
Wugen Dai, *ACT, USA*
Changhui Zhang, *ACT, USA*

A test battery is usually composed of several related test sections based on content categories. The interrelationship among different sections can be used to improve estimation accuracy and efficiency in Computer Adaptive Testing (CAT). However, using prior ability estimates to drive item selection for later test sections has not been well researched or widely applied. This paper provides an item selection algorithm that is informed by previous section scores to improve ability estimation accuracy and item selection efficiency for subsequent sections of an adaptive test battery. To evaluate this proposed algorithm, simulations were conducted under 24 conditions, following a $2 \times 2 \times 2 \times 3$ design. Four factors affecting measurement accuracy and efficiency were manipulated in administering the second section of a two-section adaptive test battery: constant versus individualized initialization of the subsequent section, maximum likelihood versus expected a posteriori estimator, loose versus tightened selection factors, and different inter-ability correlations (0.4, 0.6, and 0.8). To make the simulation study realistic, the proposed algorithm was also applied to an observed data sample. The psychometric properties and effectiveness under the various conditions were evaluated in terms of accuracy and efficiency in ability estimation, item exposure, and item usage. The simulation results suggest that the proposed algorithm achieves the best estimation accuracy and efficiency, while keeping item exposure and usage rates at acceptable levels. The observed data sample is presented to demonstrate this CAT algorithm. Choice of algorithms to optimize item selection and ability estimation for adaptive test batteries is also discussed. [29656]

17 Standardization and validation of the Inventory of Situations and Responses of Anxiety (ISRA) in the Dominican Republic

Zoilo Emilio García Batista, *Pontificia Universidad Católica Madre y Maestra, Dominican Republic*
Antonio Cano-Vindel, *Universidad Complutense de Madrid, Spain*
Kiero Guerra Peña, *Pontificia Universidad Católica Madre y Maestra, Dominican Republic*

This research had as its main objective the standardization and validation of the Inventory of Situations and Responses of Anxiety ISRA (Miguel-Tobal & Cano-Vindel, 2002) in the population of the Dominican Republic and verifies if the ISRA distinguish between healthy individuals and patients diagnosed with bronchial asthma. The total sample of 1074 subjects aged 18 to 69 years, 986 healthy subjects composed of students and the general population sample and 88 individuals with bronchial asthma, to which we applied the Inventory of Situations and Responses of Anxiety (ISRA; Miguel-Tobal & Cano-Vindel, 2002), and the Spanish version of the State-Trait Inventory (STAI, Spielberger, Edwards, Lushene, Montuori & Platzek, 1970). The data were analyzed by two-way ANOVA, showing significant differences between anxiety levels of the groups assessed. APEs confirm the existence of three sets of responses (cognitive, physiological and motor) and 3 situational areas (assessment, interpersonal, phobic and everyday life). High reliability was evidenced by Cronbach's alpha and test-retest reliability. Significant positive correlations between the ISRA and the STAI for concurrent validity were obtained. These results show that ISRA is a reliable and valid instrument to assess anxiety. [29741]

18 Comparing parametric versus nonparametric distributions of random errors in growth mixture modeling: A simulation study.

Kiero Guerra Peña, *Pontificia Universidad Católica Madre y Maestra, Dominican Republic*
Zoilo Emilio García Batista, *Pontificia Universidad Católica Madre y Maestra, Dominican Republic*

Growth Mixture Modeling (GMM) provides a more flexible approach to the analysis of longitudinal data compared to more traditional methods (Bauer & Curran, 2003). Two simulation studies are proposed to determine the performance of fit indices, likelihood ratio tests, and the consistency/replicability of parameter estimates when there are violations of the model assumptions in GMM with parametric versus nonparametric random errors. The first simulation study focuses on mixture models with growth curves having random errors that follow parametric distributions. The second explores GMM with random errors having non-parametric distributions. The main goal is to see the performance of fit indices and likelihood ratios tests under violations of the assumptions in GMM. The expected results thereof include: developing a practical guide of the benefits and risks of using the MMCC, to submit several scientific articles international peer-reviewed journals, presentations at national

and international conferences. At the institutional level this project will create a new line of research in the country and PUCMM called "Mathematical Simulations and Statistical Methodologies". in addition, it will start and foster research relations between the Pontificia Universidad Católica Madre y Maestra and the University of Missouri-Columbia. [29739]

19 The influence of social desirability response bias on self-reported engagement scores in the China College Student Survey

Fei Guo, *Tsinghua University, China*
Lin Zhao, *Tsinghua University, China*
Zhixin Lian, *Tsinghua University, China*

Quantitative empirical research in higher education relies more and more heavily on student survey data. However, the validity of the indicators/benchmarks constructed with self-reported data is often compromised by Social Desirability Response Bias (SDRB), where respondents attempt to give socially appropriate but fake responses to survey questions. Surveys on student engagement, i.e. the National Survey of Student Engagement (NSSE) in the U.S. and the China College Student Survey (CCSS), might be vulnerable to SDRB, as they ask about students' experience and behaviors on a series of "good practices" of learning. While Miller (2012) demonstrated that the influence of social desirability (SD) on NSSE's engagement benchmarks was minimal after controlling for demographic variables, a Chinese study by Jiang et al. (2013), using the CCSS 2013 data from one university, found that students' score on the SD indicator had a significant and positive correlation with their scores on the engagement benchmarks. As an extension and response to Jiang et.al's (2013) study, this paper examines the existence and magnitude of the SDRB in the CCSS survey with the national data from 2011 to 2014. It also identifies the characteristics of students with high SD scores. In addition, it proposes a way to adjust the self-reported scores of engagement indicators with the SD score following Saunders (1991). Overall, this study finds that SDRB is a non-negligible issue in college student engagement surveys in China, especially for surveys with senior students or conducted in lower-tier undergraduate colleges. [29771]

20 New item selection methods in cognitive diagnostic computerized adaptive testing: Combining item discrimination indices

Lei Guo, *Southwest University, USA*

Interest in developing Computerized Adaptive Testing (CAT) under cognitive diagnostic models has increased recently. Cognitive Diagnostic CAT (CD-CAT) attempts to classify examinees into the correct latent class profile so as to pinpoint the strengths and weaknesses of each examinee whereas CAT algorithms choose items from the item bank to achieve that goal as efficiently as possible. Most of the research in CD-CAT uses the Posterior-Weighted Kullback-Leibler (PWKL) index due to its high efficiency. The PWKL index integrated the posterior probabilities of examinees' latent class profiles into the KL information, and thus improved item selection efficiency considerably. However, the PWKL index only used examinee-based information to assess the relative importance of each latent class profile. The current study attempted to take advantage of not only the examinee-based information but also the item-based information that could be readily obtained from items. In a sense, the PWKL index should be regarded as single-source index. This paper introduced three new multiple-source item selection methods, GIDPWKL, AIDPWKL and CIDPWKL respectively, that be modified from the PWKL index by combining the item discrimination information. Two simulation studies were conducted to evaluate the new methods' efficiency against the PWKL index and Mutual Information (MI) index in the DINA model with the exposure control. The effects of different factors were investigated: the Q matrix structure (simple vs. complex), item quality (high vs. low) and test length (moderate VS. short). Simulation results indicated that: (1) In most cases, the shorter the test length was, the more advantage the three new methods would have in the fix-length test. The GIDPWKL index had the highest average attribute correct classification rate and pattern correct classification rate among the five methods. Although the new methods were superior than MI, the MI method had the best item pool utilization. (2) In most cases, the higher the item quality was, the more advantage the three new methods would have in the fix-length test. (3) The structure of the Q matrix affected the performance of different item selection methods. (4) In the variable-length test, the mean of test length across all examinees for the three new methods and MI method were all smaller than those in the PWKL method. As a whole, the performance of the GIDPWKL index was the best, and should be recommended in practice. [29578]

21 Exploring the wording effect in the Chinese version of the EPQ and NEO-FFI through item reversion

Qingke Guo, *Shandong Normal University, China*
Fei Wang, *Shandong Normal University, China*

There is an ongoing debate on the use of negatively worded (RW) items in personality and attitude assessment. Do RW items merely produce measurement bias and are they detrimental to test reliability and validity? Do RW and positively worded (SFW) items measure the same trait? These questions have not been answered clearly. In this study, the Chinese version of the EPQ and NEO-FF1 were used as research tools through item wording reversion, and were administered to participants in two week interval. The results showed that: (1) the SFW sub-scales of EPQ and NEO-FFI items are more reliable than their RW counterparts; (2) the RW sub-scales showed better construct validity than the SFW sub-scales while the factor loadings of SFW and RW items show no significant difference; (3) SFW and RW items in some sub-scales probably measured the same trait; (4) the SFW and RW effects can be removed from all sub-scales. Generally speaking, although item narration reversion has impacted item quality, both SFW and RW items should be included in personality measures. If used properly, SFW and RW worded items can tap the same construct. [29663]

22 **Cross-cultural comparison of the principal perceptions and school factors affecting students' achievement in mathematical literacy: Based on multilevel analysis of PISA 2012**
Yage Guo, *University of Nebraska - Lincoln, USA*

Student achievement is a global concern as reflected in recent large-scale standardized assessments. This study compared PISA 2012 student mathematical literacy scores with principal perceptions across four countries/regions with varying levels of student performance: Shanghai-China, the United States, Finland and Japan. Sixty-five countries participated in PISA 2012, which measured 15-year-old children's mathematical achievement and principal perceptions. The study explored the relationship of principals' perceived levels of leadership, school autonomy, and educational resources with student attainment of mathematical literacy. School variables were treated as covariates when each effect of principal leadership was interpreted. All variables were included in a multilevel model and analyzed simultaneously. The means and standard deviations of outcome variables and the explanatory and control variables for the model of the study were calculated by taking into account sampling weights, as well as plausible values for mathematical literacy scores. SAS PROC MIXED was used to fit hierarchical linear models for the study. The findings indicated that: with students' background controlled, the effect of school educational resources on students' mathematical literacy demonstrated some cultural differences among the four countries. Specifically, class size had a significantly positive effect on students' mathematical literacy in Finland and Japan. There was a negative relationship between student achievement and lack of educational resources. Social, economic, and cultural status showed a positive relationship with mathematical literacy under each of the four different cultural contexts. Results also indicated that students are likely to achieve better if principals perceive that there are no shortages of personnel and equipment. [29761]

23 **A study of the impact of student individual factors and school factors on academic achievement**
Jilin Han, *Beijing Normal University, China*

The academic achievement of students is one of the basic indicators of the quality of school education. Research on factors affecting the academic achievement of students can provide targeted guidance to educational practice. This research focused on the influence of student factors and school factors on boarding students' academic performance using hierarchical linear modeling. The results showed that there were significant differences in academic achievement of students from different schools. School-level factors explained large proportions of the total variation of the scores, but student-level factors explained more. Student background factors, such as gender, predicted academic achievement significantly; and students' individual perception of teachers' responsibility and love, and perception of class order situation had significant predictive effect on academic achievement. Besides, after controlling for all independent variables at the student level, school resources (shortage of school teachers, teaching resource abundance degree) significantly predicted academic achievement of students. In addition, the cross-level interaction between gender and school resources factors was not significant. [29848]

24 **Performance of fit measures in mis-specified models.**
Woori Han, *Sungshin Women's University, South Korea*
Yumi Yi, *South Korea*

The Latent Growth Curve Model (LGCM) is popular for analyzing longitudinal data. This study investigated the performance of fit measures in mis-specified LGCMs. The study used a factorial design of 2(df) by 6(sample sizes) by 4(mean of quadratic). Samples size conditions included the six levels of 50, 100, 200, 400, 600 and 1000. The last condition was the mean values of the quadratic

factor. We chose relatively small values (i.e., 0.1, 0.01, 0.001 and 0.0001) to create a mis-specified model. Monte Carlo Simulations were employed for the generation of longitudinal responses following a quadratic LGCM. Except for the values of the quadratic growth factor, the parameter values were those used by Kenny et al. (2014). All models were estimated by using Mplus. None of the fit indices should be trusted when the model has a small df. Specifically, when the df and the quadratic value are small, RMSEA, CFI and TLI values indicated that mis-specified models fit well. Meanwhile, RMSEA was sensitively influenced by df and sample size. The larger sample size was, the smaller RMSEA was in mis-specified models having small df. On the other hand, the value of RMSEA was large in mis-specified models having relatively large df regardless of sample size. In addition, the RMSEA varied considerably according to df while the values of CFI and TLI were relatively stable. The significance and limitations of the result are discussed. [29929]

25 Why students stay: An exploratory study of factors impacting postsecondary students' institutional commitment

Ye He, *The University of North Carolina at Greensboro, USA*
Bryant Hutson, *The University of North Carolina at Greensboro, USA*

Student retention is one of the most important considerations for postsecondary institutions, especially in small private colleges where resources may be limited. Research on this topic tends to focus on examining the causes of retention problems, while institutional strengths and practices that lead to students' decisions to stay are often neglected. To better leverage institutional strengths for contextualized retention strategies, this study was designed to explore factors that are associated with students' decision to stay enrolled in small, private colleges through the development and use of a survey instrument, Why Student Stay (WSS). A mixed-methods sequential design was employed in this study (Creswell & Plano Clark, 2011). Participants included administrators, faculty, staff, and first-year students from six small, private colleges. The study was carried out in two phases: 1) WSS construct and item development based on literature review, expert feedback, and stakeholder interviews; and 2) institution profile construction based on regression model building. Qualitative data were first collected for instrument development and construct identification. During the second phase, logistic regression was used to explore WSS factors associated with students' decision to stay, while controlling for other independent variables including students' ethnicity, gender, age, family income level, predicted academic performance, and academic outcomes. The findings of this study led to the identification of six significant factors associated with student retention. The process of institutional retention profile construction based on these factors empowered stakeholders to build institutional-specific retention models that leveraged existing resources and successes. [29524]

26 A multivariate measure of Chinese family involvement in early childhood education

Xiaoting Huang, *Peking University, China*
Kexin Guan, *Peking University, China*

Family involvement has been found to be an important factor in children's development. This study developed and evaluated a Chinese Family Involvement Survey using Wilson's four-building-blocks approach. Guided by theory and adapted to the Chinese society, the survey consists of three involvement constructs: nutrition and health, family educational activity, and educational expenses. Demographic and survey question data were collected from 588 parents. A multidimensional Rasch model was applied to analyse the data. Results suggested that the reliability of the three sub-scales ranged from 0.61 to 0.63. The overall reliability of the survey was as high as 0.9. No significant difference was found between girls and boys in nutrition and health, or family educational activities. However, parents reported spending more on girls' early education. Rural families were shown to have lower educational expenses, and parents engage in less family educational activities. Single parent households were also found to be less involved in educational activities. In addition, results indicated that Chinese parents generally care more about children's nutrition and health than other child rearing aspects. [29625]

27 The strategy of science popularization in college under the "Otaku" culture dilemma

Gao Jian, *China Research Institute for Science Popularization, China*

Under the backdrop of the information age, the influence of "Otaku" culture among Chinese university and college students has recently grown. Therefore, both awareness and analysis of the "Otaku" phenomenon on campuses are needed. In the composite perspective of science and technology communication worker and educator, the mental, behavioral, and cognitive situation of "Otaku" should be defined and distinguished through literature

collection and research, combined with case study and practice. Viewpoints and suggestions in improving scientific and technological information communication in the "Otaku" cultural atmosphere should be put forward. [29474]

28 Using the Rasch model to develop a questionnaire to measure students' chemistry learning motivation level

Yanxia Jiang, *Shandong Normal University, China*
Lei Wang, *Beijing Normal University, China*

In this study we used a Rasch model to develop a questionnaire which measures students' Chemistry Learning Motivation Level (CLML). Based on motivation theory, we construct a theoretical model of chemistry learning motivation level which consist five motivation levels: negative, passive, active, positive and creative. In total, 354 junior high school students from China, varying in grades, sex, motivation levels and achievements, were selected by random sampling to respond to the questionnaire. Analysis, statistical treatment, and modification of the questionnaire were carried out, and minor events were further elaborated, which resulted in a questionnaire with 23 items. Cronbach's alpha for the entire questionnaire was 0.904. The IMNSQ and OMNSQ for the entire questionnaire were 1.00 (ZSTD=-0.2) and 1.02 (ZSTD=-0.1). The person reliability and item reliability were 0.86 and 1.00. The person separation and item separation were 2.49 and 16.58. A unidimensionality test showed that the raw variance explained by the measures was 61.3%. The other important indicators, such as the Wright map, item analysis, standardized residual variance, and bubble plots were ideal. High level motivation students and low level motivation students showed a significant difference on their CLML scores. Findings of the study confirmed the validity and reliability of the CLML questionnaire. [29638]

29 Cross-year test equating based on item response theory

Xiong Jianhua, *Jiangxi Normal University, China*
Ding Shuliang, *Jiangxi Normal University, China*

To monitor and assess the quality of education, a comparison of students' academic ability across different years or regions is usually needed. This is an example of a test equating process. Test equating requires either the same examinees to take two different test forms or two examinee groups to take the same test. In China, in order to measure a student's academic ability, the test needs a wide coverage, an adequate item bank, and to be offered regularly, such as the senior high school entrance examination and the college entrance examination. However, for these tests, the test forms are published, different regions use different test forms, and a majority of the examinees take only one test. This study uses a Monte Carlo simulation. First, students' scores in different years and in different regions were simulated according to a design scheme, then the scores were linked by equipercentile equating, and finally score changes were analyzed. Under item response theory and using a parameter estimation program, the item parameters of each test and the ability parameters of each student are estimated such that the ability and item difficulty parameters are on the same scale. By calculating the test information, this study can evaluate the quality of each test and find the most suitable group for the test, which is valuable. [29558]

30 Estimating classification accuracy indices for cognitive diagnostic testing under the DINA model

Jiaxuan Du, *Beijing Normal University, China*

In Cognitive Diagnostic Assessment (CDA), the attribute classification-based decisions are made after calibration, while the validity of classification is often not reported in diagnostic score reporting. Classification accuracy indices (Cui, et al., 2012; Wang, et al., 2014) were used as important indicators of the validity of classification results to evaluate the accuracy of classification results, while the Wang method not only includes the results at the whole-pattern level, but also develops attribute-level classification accuracy indices for CDA. The study investigated attribute-level and pattern-level classification accuracy indices based on the Lee method and Wang method in CDA. Methods were compared using a simulation study and an empirical study. The deterministic input noisy "and" gate model (DINA) was used to evaluate the performance. According to the simulation results, the classification accuracy indices performed well with simulated diagnostic tests in that their values matched the simulated correct classification rates closely in the middle and large sample conditions. The number of attributes and the values of the guessing and slipping parameters can also influence the classification accuracy indices and classification rates. The results in the real diagnostic data also support the conclusion that the newly developed indices can be used as validity indicators in CDA. [29572]

31 The theory and practice of psychological "process measurement" research

Zhang Jinhua, *Guang'anmen Hospital, China*
Weidong Wang, *Guang'anmen Hospital, China*
Xueyu Lü, *Guang'anmen Hospital, China*
Fan Feng, *Guang'anmen Hospital, China*

In this paper, through the personality development "process measurement", we develop an understanding of personality development stage and the measurement of personality development level. On the basis of psychological theory, and ideas of Traditional Chinese Medicine (TCM), the holistic view, dialectic view, the concept of dynamics etc., we integrate personality development and TCM ideas using literature analysis, qualitative research, as well as brain storm of experts. This information is then used to extract the theoretical basis of abnormal personality development theory, to build the architecture of the questionnaire and enact items. The age of questionnaire's character is then graded. Results: (1) Formed abnormal personality development theory (while taking into consideration that there are offset and missing in the development of the personality), as the development of different ages have different characteristics, such as the case where the individual at the age of 25 years old would now be stable in the personality development. (2) Determined the basic structure of the "process measurement" questionnaire, where the preliminary questionnaire was divided into childhood, adolescence, adulthood, where childhood is 3 years old to preschool, adolescence is from 7 years to 17 years old, and adulthood is from 18 years old to 25 years old. Conclusion: (1) The TCM view, holistic view, dialectic view, and concept of dynamics play a huge role; (2) The personality development evaluation of "process measurement" is an important part of measurement, as the study demonstrated that process measurement and evaluation of personality development provides an important basis for clinical psychological treatment and guidance. [29956]

32 An application of the Thurstonian item response theory model to personality data

Kentaro Kato, *Benesse Educational Research and Development Institute, Japan*

The Thurstonian Item Response Theory (IRT) model is a multidimensional IRT model for Forced-Choice (FC) data, which is intended to reduce the effect of response faking. In this study, a 69-items personality questionnaire was administered to two groups of college students in a non-faking-invoking situation. The first group of examinees responded to the questionnaire in the five-point Rating Scale (RS) format. Exploratory and confirmatory factor analysis revealed that eight factors well explained the relationship between items. The second group responded to the same questionnaire in the FC format, in which 23 blocks of triplet comparisons were made. An eight-dimensional Thurstonian IRT model was applied to the FC data to estimate the model parameters. Equivalence of RS and FC subscale scores was examined for common examinees in the first and second groups. [29599]

33 Developing a report card management program for high school

Keonseob Kim, *Korea Institute for Curriculum and Evaluation, South Korea*
Insuk Kim, *Korea Institute for Curriculum and Evaluation, South Korea*
Youn-Joon Joung, *Korea Institute for Curriculum and Evaluation, South Korea*
HeeKyoung Kim, *Korea Institute for Curriculum and Evaluation, South Korea*

According to the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) results, Korean students show high academic performance with a low index of affective characteristics, such as interest, self-efficacy, and value perception. Moreover, a recent world survey on people's happiness reported that the level of Korean students' happiness is relatively low. This result is presumably caused by the Korean education system, which mainly focuses on the college entrance examination. In this regard, President Park Keun-Hye's regime has shifted its focus on educational state affairs toward 'education for happiness,' with policies that encourage students' dreams and talents. 'Education for happiness' is a policy that breaks from traditional knowledge-based learning, and shifts to student-centered learning where students voluntarily participate in learning and plan their future. In order to adapt this education policy to the classroom, a customized education suited to each individual student is needed, which requires a break from the traditional type of report card system that simply provides scores, grades, and ranks to the students after any assessment. Therefore, investigating ways to introduce a practical report card system that indicates students' strengths and weaknesses based on their performance in a direct feedback form has much practical value. The purpose of this study is to develop and introduce a practical high school report card system that provides customized rating information to students and the ways in which it provides achievement levels in each specific academic element after each term exam at Korean high schools. [29623]

34 Assessing psychometric properties of learning in organizational performance at the organizational level

Kyoungshin Kim, *University of Georgia, USA*
Karen E. Watkins, *University of Georgia, USA*
Zhenqiu (Laura) Lu, *University of Georgia, USA*

Developing sophisticated diagnostic measures of organizational performance can be useful for improving organizational effectiveness. Although many studies have developed strong measures of individual performance, we do not yet have equally strong measures of organizational performance because of the complex nature of organizations. Developing measures of organizational performance impacted by learning and development activities could be more challenging. This study explored psychometric properties of learning in organizational financial performance and knowledge performance by adapting the dimensions of the Learning Organization Questionnaire developed by Watkins and Marsick. Factor Analysis (FA) and Structural Equation Modeling (SEM) have been widely used in the social and behavioral sciences. This study combined second order FA with SEM and applied it to the measure of organizational performance and to examine the relationships between learning organization, knowledge performance, and financial performance. The results showed there were both direct and indirect effects of the learning activities in organizations on organizational financial performance and knowledge performance. Knowledge performance partially mediated the effect of learning activities in organizations on organizational financial performance. Model fit indices, model modification, and related implications were also discussed. [29920]

35 Interval estimation for true raw and scale scores under the compound binomial error model

Kyung Yong Kim, *University of Iowa, USA*
Won-Chan Lee, *University of Iowa, USA*

Measurement errors for stratified domains can be modeled using a compound binomial distribution when an examinee's observed scores over repeated measurements, conditional on true proportion-correct score for each domain, follow a binomial distribution. Under the compound binomial error model with two stratified domains, the purpose of the present study is to introduce the Wald confidence interval, the score confidence interval, and the likelihood ratio confidence interval for true score; to present the Wald confidence interval and an endpoints conversion method for true scale score; and to compare the interval estimation procedures for both true raw and scale scores in terms of their coverage probabilities and interval widths using a simulation study. Since confidence intervals for linearly transformed scale scores have the same coverage probabilities as their corresponding intervals for raw scores, only nonlinearly transformed scale

scores will be considered in this study. Furthermore, confidence intervals constructed under the compound binomial error model are compared to those constructed under the binomial error model, which ignores stratification. [29552]

36 Comparison of random effects and fixed effects item response models for detection of a multilevel treatment effect

Meereem Kim, *University of Georgia, USA*
Zhenqiu (Laura) Lu, *University of Georgia, USA*
Allan S. Cohen, *University of Georgia, USA*

A multilevel random effect IRT model is regarded as a powerful analytic tool because such a model can estimate the randomness of item parameters which are assumed to be fixed for the traditional multilevel IRT model. However, there is little research comparing the performance of these models. The purpose of this study is to compare the results from multilevel random effects IRT models and traditional multilevel fixed effects IRT models. Simulation studies were conducted to investigate the impact on both the item parameter and the ability parameter estimates of using a multilevel Rasch model, in which items are nested within individuals and individuals are nested within organizations. Sample sizes at all three levels were varied. Specifically, the simulation included 48 different conditions: 2 multilevel models (multilevel fixed effect Rasch model and multilevel random effect Rasch model), 3 sample sizes for organization (20, 40 and 80 schools), 3 sample sizes for individual (500, 1,000 and 2,000 examinees), and 3 sample sizes for item (10, 20, and 40 items). Thirty replications were simulated for each condition. Model parameters were estimated for both multilevel IRT models and their performance was examined in terms of parameter recovery. Initial results from a pilot study showed that the random effect Rasch model performed well compared to the fixed model. [29910]

37 Generalizability theory approach for estimating reliability of testlet-based tests with balanced and unbalanced data

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

This study is aimed to investigate the measurement error and reliability of test scores composed of testlet with balanced and unbalanced data structures. Lee (2003) differentiated three types of measurement designs: 'item' based design ($p \times I$ design), 'testlet' based design ($p \times H$

design), and 'item(s) nested in testlet' design ($p \times (I : H)$ design). This study focused on the magnitude of measurement errors when applying the three different designs with simulated data sets. Two different types of data are simulated – balanced and unbalanced. Balanced data structures mean that the same numbers of items are allocated into testlets. Unbalanced data structures cannot satisfy this requirement. This study will explain the reason why error of measurement occurs when using an 'item' based design and a 'testlet' based design by mathematical formulas. [29682]

38 Obtaining a representative sample when the distribution of elements is not known

Sewon Kim, *Michigan State University, USA*
Unhee Ju, *Michigan State University, USA*

Stratified two-stage sampling is frequently used in the field of educational research, especially when a list of elements does not exist but a list of clusters can be created. However, this method is problematic due to the high rejection rate of participation and to not reaching the desired sample size. International studies have generally faced these challenges, due to different educational environments across countries. To grapple with these problems, we propose a stratified two-stage sequential sampling approach, based on modified adaptive cluster sampling with stratification in the field of ecology (Christman, 2003). The proposed sampling is expected to outperform stratified two-stage sampling under the uncertainty of the distribution of elements and a dynamic population, in terms of cost effectiveness and precision of estimation. Thus, in this study, stratified two-stage sequential sampling is demonstrated using the data available from the Wisconsin Information System for Education, and its efficiency over stratified two-stage sampling is evaluated with respect to cost and the precision of estimates in a simulation study under different conditions: (a) sample size (50, 100, 500, and 1,000), and (b) participation rate (0.2, 0.5, and 1.0). Results will be presented to show the performance of the two sampling approaches under different simulation scenarios, and they will possibly provide guidance when the distribution of elements is not known, or when there is low probability of participation in educational studies including international studies. [29647]

39 Estimating categorical marginal models: Comparing ML and GEE using simulations

Renske E. Kuijpers, *Leiden University, The Netherlands*
Mark J. de Rooij, *Leiden University, The Netherlands*

Categorical marginal models are flexible models for analyzing dependent categorical data without making specific assumptions about the nature of these dependencies. For estimating categorical marginal models, different estimation methods can be used. Kuijpers, Bergsma, Van der Ark & Croon (2014) explored to what extent maximum likelihood (ML) estimation and generalized estimating equations (GEE, Liang & Zeger, 1986) are appropriate for investigating and testing different types of research questions. Results showed that both methods have advantages and disadvantages. For instance, GEE has problems with efficiency and does not readily provide goodness-of-fit statistics, whereas ML becomes problematic for large sets of items. However, they only investigated the methods for three very basic research cases. Furthermore, they did not investigate how the estimation methods handle missing data. In this new study, we simulated data in order to create more advanced and realistic research situations, and included missing data. GEE and ML estimation are compared based on the bias of the model parameters' estimates and standard errors, on the coverage of the corresponding confidence intervals, and on their convergence rates. Furthermore, the methods are compared with respect to handling missing data. [29724]

40 Comparing unidimensional and multidimensional IRT models in different sample conditions

Gökhan Kumlu, *Hacettepe University, Turkey*
Hülya Kelecioğlu, *Hacettepe University, Turkey*
Gulfem Dilek Yurttaş, *Gazi University, Turkey*

Unidimensionality is the assumption which expresses that all the test items which affect the test score measure the same latent trait and the test performance of the examinee results from only this latent trait (Pomplun, 1988). Multidimensional IRT is adaptation of unidimensional IRT models to multidimensionality (Ackerman, Gierl & Walker, 2003). The goal of this study is to compare the unidimensional and multidimensional IRT models at different sample sizes and to determine which model has a higher measurement accuracy. The data were obtained from the results of the 2012 High School Placement Test applied to 8th graders in Turkey, and a 40-item data set was formed from Maths and Science questions of the test. To analyze the effect of sample size, simple random samples of 500, 1000, 2000, and 4000 examinees were chosen from the total population of 258,830 examinees. To compare the models, residual and error variance, SSR and RMSR, and Tanaka index of goodness of fit were calculated under the two models. In cases where unidimensionalty assumption could not be completely met,

the data show better adjustment with the multidimensional IRT model. Additionally, the multidimensional IRT model gives improved model-data fit if the sample size is large. Within the multidimensional model, item parameters were found to be less erroneous and have a greater measurement accuracy compared to the unidimensional model. For multidimensional model, it was found that as the sample size increases, the error decreases and the measurement accuracy improves. [29592]

41 Bayesian model averaging for near-equivalent path models

Keke Lai, *University of California, Merced, USA*
Yuzhu (June) Yang, *University of California, Merced, USA*
Sarah Depaoli, *University of California, Merced, USA*

Model selection for path analysis involves at least two principles. First, a good model should be parsimonious, so that it is easy to interpret and generalize. Second, researchers should consider other models that fit the data similarly well as does the proposed model, as those models represent equally plausible explanations. Unfortunately, these principles are often violated in the practice of path analysis. On the one hand, a parsimonious path model often does not exhibit adequate model fit, and thus researchers may find the model difficult to justify and resort to a highly saturated model for better fit. On the other hand, a highly saturated model often entails many near-equivalent models that can fit the data similarly well, and thus those competing models weaken the credibility of the proposed model. To avoid this dilemma, instead of looking for a "best" model, we proposed a Bayesian method to incorporate in the analysis all the possible models and the uncertainty of each model. In particular, each model has a prior probability that is positively linked to the model's parsimony. Given the data, we can obtain the posterior distribution for the model parameters of interest. This posterior distribution is an average based on the posterior distributions under each of the possible models considered, weighted by their posterior model probability. In so doing, we account for the uncertainty in model selection and results can better generalize to similar situations. In this presentation we will also demonstrate our approach with a real-data example. [29743]

42 Effectiveness of three methods for a just-identified mean structure of MTMM data using structural equation modeling

Mina Lee, *Sungkyunkwan University, South Korea*
Soonmook Lee, *Sungkyunkwan University, South Korea*

Interpretation of mean structure analysis on Multi-Trait-Multi-Method (MTMM) data has not been attempted in practice because it has not been viewed meaningful to impose a mean structure on a single set of data such as MTMM data. However, scores obtained by different methods in MTMM data can be viewed as data of different groups, thus it is reasonable to treat MTMM data as a special type of multi-group data. In order to make the mean structure of MTMM data interpretable substantively, the mean structure should be just-identified. The mean structure of MTMM data can be just-identified by adopting a reference-group method, marker-variable method, and effects-coding method delineated by Little, Slegers, & Card (2006). We examined the performance of the three methods. The results showed that effects-coding method performed better than the marker-variable method, and the marker-variable method performed better than referenced-group method in terms of estimation accuracy. Also we showed interpretation according to the effects-coding method applied to real data. [29799]

43 Using asymmetric ICCs to account for the effects of guessing on multiple-choice test items

Sora Lee, *University of Wisconsin - Madison, USA*
Daniel Bolt, *University of Wisconsin - Madison, USA*

Many applications of item response theory to multiple-choice item response data make use of the three-parameter logistic (3PL) model to account for the effects of guessing. An alternative approach considers a model that introduces asymmetric ICCs, such as the residual heteroscedasticity (RH) model considered by Molenaar (2014). Using several item response datasets, we empirically compare the fit of the 3PL and RH models to actual response data. We consider some differences between the two approaches with respect to how guessing effects are accounted for, as well as some potential advantages of the RH model in applied IRT settings. [29932]

44 Parameterizing the general diagnostic model to examine differences in attribute distributions between multiple groups

Young-Sun Lee, *Columbia University, USA*
Jung Yeon Park, *Columbia University, USA*
Nayeon Yoo, *Columbia University, USA*

Cognitive Diagnosis Models (CDMs) are a type of latent class model used to diagnose and categorize each student with a fine-grained skill pattern he or she possesses. The general class of diagnostic model, so called 'GDMs' (von Davier, 2005; von Davier & Yamamoto, 2004) is a type

of CDMs but was developed with the goal of maintaining similarities to previous approaches using ideas from IRT, log-linear, and latent class analysis models. One of the advantages of the GDM lies in its applicability to attribute models for polytomous item responses and to attributes with more than two proficiency levels. For the purpose of comparing the skill distributions of students across multiple groups (e.g., country, gender, etc.), the model can be also extended to 'multiple group' GDMs. In this study, we compare estimated attribute distributions of students in the United States, Republic of Korea, and Russian Federation using 2011 TIMSS Grade 8 Mathematics database. To detect if there are differences in attribute distributions among the groups, we employed two statistical tests – the likelihood ratio test and the Wald test. In order to consider the multi-stage cluster sampling technique used for the TIMSS data, we used Jackknife replicated weights to approximate covariance matrices of the prevalence estimates for the two tested aforementioned. We present the performance of the methods via a simulation study with different test lengths and sample sizes. [29927]

45 The development of online Chinese collaborative problem solving tests in science

Cheng-Hsuan Li, *National Taichung University of Education, Taiwan*
Kai-Chih Pai, *National Taichung University of Education, Taiwan*
Bor-Chen Kuo, *National Taichung University of Education, Taiwan*
Yinung Lin, *National Taichung University of Education, Taiwan*
Zhi-Yong Liu, *National Taichung University of Education, Taiwan*

Collaborative Problem Solving (CPS) is a critical "21st century skill" in education. In addition, OECD PISA assesses 15 collaborative problem skills for communications, interactions, and collaborations of test-takers with computer agents through the use of multiple choice items in 2015. However, the released test items in the PISA 2015 draft CPS framework were in English, and only multiple choice items were considered. In this study, with reference to the 15 skills in the framework of the draft, two online Chinese CPS tests in science for Taiwan students, Water Purification and Water Jar problems, consisting of both multiple choice items and constructed response items were developed. In Water Purification, a test-taker would need to collaborate with two agents, a high collaboratively-orientated agent and a low collaboratively-orientated agent, to design a simple water

purifier in school competition. The high collaborative-oriented agent always reveals information about the task. On the other hand, the low collaborative-oriented agent sometimes interrupts and disagrees with others. In the other test, Water Jar Problem, a test-taker and an agent would have to balance an object using compound pulley and two bottles whose masses are not equal to the object. The test-taker is able to do some experiments using the computerized fixed pulley and moving pulley. At the same time, the agent records the progress made in the experiments. Moreover, they have to communicate with each other to design an appropriate compound pulley, so as to obtain suitable liters of water. Some results will be shown in this study. [29862]

46 The application of Thursonian IRT models in the development of anti-fake forced-choice questionnaires

Hui Li, *Beijing Normal University, China*
Hongyun Liu, *Beijing Normal University, China*

Forced-choice questionnaires have been controversial because traditional scoring methods result in ipsative data, whereby all individuals have a common total test score. For addressing this problem, the Thurstonian IRT model has proved truly helpful. However, previous studies did not consider social desirability when pairing items and thus rendered the questionnaires vulnerable to fake choices. This study is a simulation study in which we matched social desirability when pairing items and compared the applications of the Thurstonian IRT model under different conditions. Forty eight conditions were examined by crossing the following factors: (a) number of measured traits (2 or 5); (b) number of items per trait (10 or 20); (c) percentage of positive-negative item pairs (0%, 20% or 40%); (d) percentage of unidimential pairs (0% or 20%); (e) trait correlations (0.2 or 0.7). The results are as follows: (1) the latent trait scores and item parameters are recovered more accurately when the number of traits is larger; (2) a larger number of items per trait indicates better model estimates and more stability; (3) larger percentage of positive-negative item pairs leads to better model estimates; (4) the impact of the percentage of unidimential pairs is not significant; and (5) increased correlation among traits worsens the recovery of latent trait scores and model convergence but improves the precision of parameter estimation, so the impact of it should be further examined. We hope the current research could provide proper guidance for designing forced-choice questionnaires, especially tests that can resist faking in the future. [29696]

47 The effects of mathematics self-efficacy and school climate to mathematics achievement: A multilevel analysis based on PISA 2012

Hua-Xia Li, *Beijing Normal University, China*
Tao Yang, *Beijing Normal University, China*
Danhui Zhang, *Beijing Normal University, China*
Qin Yi, *Beijing Normal University, China*
Yi-Bo Wang, *Beijing Normal University, China*

Both mathematics self-efficacy (individual-level) and school climate (school-level) play important roles in mathematics achievement. A hierarchical linear model was applied to analyze the effects of students' mathematics self-efficacy and their school climate based on Shanghai PISA 2012 data. The results indicate that: (1) when we control for individual-level gender, Economic, Social and Cultural Status (ESCS), sense of belonging to school, and school-level school ownership and mathematics teacher-student ratio, mathematics self-efficacy with its covariates such as mathematics self-concept, mathematics interest, and attribution to failure of mathematics significantly effect mathematics performance. (2) Mathematics motivation, mathematics anxiety and mathematics behavior do not produce a significant effect on mathematics achievement. (3) School-level factors, student-related factors affecting school climate and teacher-related factors affecting school climate are significant predictors. (4) Teacher focus, teacher participation/autonomy, teacher morale, and shortage of teaching staff do not directly predict mathematical performance, but the two variables teacher morale and shortage of teaching staff generate indirect effects on mathematics achievement via student-related and teacher-related factors affecting school climate. In teaching practice, it is very important to strengthen students' mathematics self-efficacy and to strengthen school climate. [29865]

48 Group types of job stress for university teachers in China

Xiaoyue Li, *Shandong Normal University, China*
Fengqiang Gao, *Shandong Normal University, China*
Peng Wang, *Shandong Normal University, China*
Fang Wang, *Shandong Normal University, China*
Yu Tian, *Shandong Normal University, China*

The objective of this study was to group the types of job stress for university teachers in China. Through layered random sampling, 1,988 teachers from 22 colleges and universities from all over the country were measured by the indigenous Scale of Job Stress for University Teachers and other tools. A K-Means clustering method was used for classification of university teachers. According to the characteristics of job stress, Chinese university teachers can be classified into six significantly heterogeneous types – relatively high stress, task-related stress, relatively low stress, interpersonal-related stress, extremely low stress and extremely high stress. There are significant differences among different types. Scores on job burnout and job satisfaction of different job stress types for university teachers show different features. The distributions of types of job stress are significantly different among genders, school types, and turnover intention groups. In result, job stress of university teachers can be divided into different types and this classification model can make pointed references for stress management and psychological aid for university teachers. [29787]

49 A polytomous extension of the Hamming distance discrimination method

Yujun Li, *Jiangxi Normal University, China*
Zhaosheng Luo, *Jiangxi Normal University, China*
Xiaofeng Yu, *Jiangxi Normal University, China*
Chunlei Gao, *Jiangxi Normal University, China*
Yafeng Peng, *Jiangxi Normal University, China*

Cognitive Diagnosis Assessment (CDA) aims to gain more information about examinees' cognitive strengths and weaknesses at the skill, attribute, or competence level. Although there has been much research focusing on CDA, most research has been only for dichotomous data. This has limited the application and development of CDA. In practical applications, test designers show a preference for polytomous items. From this perspective, more research about how to make cognitive diagnosis models appropriate for polytomous responses is required. The Hamming Distance Discrimination (HDD; Luo, Li, Yu, Gao, & Peng, 2015) has been shown to have good performance using simulation data. The HDD uses the Hamming Distance (HD) to measure the distance between an examinee's Observed Response Pattern (ORP) and an Expected Response Pattern (ERP). In this paper, the HDD was extended to the analysis of polytomous response in cognitive diagnostic assessment. The similarities between the ORP and ERP for the polytomous response situation are measured by the HD, and the attribute patterns can be recognized via the relationship between attribute patterns and IRPs. The design and the results of a simulation study are presented, and the classification accuracy of the HDD for polytomous data under different conditions is investigated. The results of this study demonstrated that for polytomous data, the HDD was easy to be understand and implement. Moreover, the HDD had a high classification rate. [29691]

50 Developing a scale for interest of elementary students in math learning

Miaoyun Li, *Southwest University, Chongqing, China*
Naiqing Song, *Southwest University, Chongqing, China*

Interest makes an important contribution to the math performance of students in primary schools. Most of the literature on interest in math learning mostly focuses on teaching strategies, and few studies consider the measurement of learning interest. To provide theoretical evidence and practice instructions for promoting interest in math learning for students in primary schools, a scale was developed to quantify interest of elementary students in math learning. The research included two main parts. The first was to develop an initial item collection based on related literature and interviews, and then to explore the structure of the scale via the exploratory factor analysis using SPSS21.0. After deleting some invalid items, a sample of 300 primary schools students was analyzed. A five factor structure was obtained and the five factors were named as understanding, pleasure, engagement, exploration, application, respectively. The five-factor structure was further validated in the second study via confirmatory factor analysis in Mplus 7.0. The results showed that the structure fits well with the data from another sample of 300 students. At the same time, the instrument has acceptable reliabilities for all subscales. In addition, gender invariance and grade invariance were examined. Finally, the improvement and application of this instrument were discussed. [29710]

51 Factor analyses of the Chinese college entrance examination

Kai-Li Liang, *Beijing Normal University, China*
Xian-Wei Liu, *Beijing Normal University, China*
Hong-Bo Wen, *Beijing Normal University, China*

The Chinese College Entrance Examination is an important means to examine students' language proficiency. Studies have focused on reliability and validity of language papers, but few studies have investigated the structure of language proficiency. The syllabus pointed out that the Chinese test measured three subscales, including language and knowledge, literary knowledge and reading comprehension of ancient poetry and modern culture. The question remains, however, as to whether the items on the various forms of the Chinese test behave as intended given the proposed interpretation of the three subscales quoted above. Is the intended structure of the Chinese test actually the observed structure? One purpose of this study, then, was to explore the factorial structure of the current Chinese test. The other purpose was to confirm the results and examine competing models. The results showed that the Chinese language proficiency test was single-dimensional; even when items from language and knowledge, literary knowledge, reading comprehension of ancient poetry, and modern culture are specified with separate factors, these factors appeared very highly related. In addition, there was a difficulty factor for difficult items. [29837]

52 Joint assessment of closeness between factor analysis and principal component analysis, and consistency

Lu Liang, *University of Hawaii at Manoa, USA*
Kentaro Hayashi, *University of Hawaii at Manoa, USA*
Ke-Hai Yuan, *University of Notre Dame, USA*

Guttman (1956) suggested that the population loadings of Factor Analysis (FA) and those of Principal Component Analysis (PCA) approach each other if the number of variables p goes to infinity while the ratio of number of factors m to the number of variables p goes to zero. It implies that as p increases, PCA gives a closer approximation to FA. Because the computation for the estimates of PCA loadings is simpler than the computation for the estimates of FA loadings, we can use PCA as an approximation for FA when p is large. However, another side of the same coin is that as p increases relative to the fixed sample size N, the estimates of PCA or FA loadings might not be close enough to their population loadings. That is, non-consistency might become an issue as p increases. Therefore, we need to simultaneously consider the closeness between the estimated FA loadings and the estimated PCA loadings as well as the closeness between the estimated PCA loadings and the population FA/PCA loadings. We study how to achieve the two goals simultaneously: (i) to make it close between the estimated FA loadings and the estimated PCA loadings, and (ii) to make it close between the estimated FA/PCA loadings and the population FA/PCA loadings. We study this issue under the high-dimensional conditions, e.g., with large p. To deal with the high-dimensionality, we employ the ridge estimator as our regularization method proposed by Yuan & Chan (2008). [29551]

53 Development of item selection methods in cognitive diagnostic computerized adaptive testing

Chuan-Ju Lin, *National University of Tainan, Taiwan*

Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) has received much attention recently. The

primary purpose of this research is to develop a new item-selection method simultaneously balancing attribute coverage and item pool usage, while maintaining acceptable estimation precision in CD-CAT. The secondary goal in this research is to derive an item selection index based on mutual information (MI) for CD-CAT. A simulation study will be conducted to examine the effectiveness of the new methods by comparing them with previous item-selection methods for CDMs in psychometric precision and non-psychometric features of the resultant CD-CATs (i.e., degree of attribute balance and degree of exposure balance). Variables examined that might influence CD-CAT characteristics included (a) test length, (b) number of attributes, (c) item pool size, (d) correlational structure of the abilities (or attributes), and (e) the proportion of mastery for each attribute. [29689]

54 The effects of emotional connections in convalescent heroin addicts

Jie Liu, *Southwest University, Chongqing, China*
Ning Zhou, *Yunnan Normal University, China*
Haiyan Zhu, *Yunnan Normal University, China*
Jinfu Zhang, *Southwest University, Chongqing, China*

The theoretical models relevant to drug addiction have explained the key factors and mechanisms of drug initiation, as well as the maintenance of drug use, from various angles. The emotional pathway related to addictive behavior and relapse is definitely confirmed in these theories and previous studies. However, the efficacy of this kind of emotional pathway has not been adequately defined. In the present research, the Positive Affect and Negative Affect Scale (PANAS) and Heroin Craving Questionnaire (HCQ), accompanied by an experiment conducted with heroin related stimulus, are administrated in convalescent heroin addicts under methadone maintenance treatment to explore the relative strength of the connection between positive emotion cues and negative emotion cues. The main findings are as follows: (a) There is a significant correlation between negative affect scores and heroin craving scores, and negative emotion status properly predicts heroin craving in addicts ($F = 6.3334$, $p < .05$); (b) Emotional connection with heroin related stimulus exists both in positive emotion cues and negative emotion cues, and the effects associated with negative emotion cues are significantly stronger than that of positive emotion cues ($t = 2.705$, $p < .05$); (c) The results from the experiment are in accordance with the findings based on the measure, suggesting that PANAS and HCQ are useful for exploring emotional connection among heroin addicts. [29826]

55 The influence of the distribution of task difficulty on the method of improving test reliability in nested designs

Xian-Wei Liu, *Beijing Normal University, China*
Hong-Bo Wen, *Beijing Normal University, China*
Pei Yin, *Beijing Normal University, China*
Kai-Li Liang, *Beijing Normal University, China*

For improving the reliability of nested design (i:t), some researchers think that the effective way is to increase the number of tasks (T), but others believe that it is better to add items to each task. Analysis of previous studies finds the difference between former tasks' difficulty and the homogeneity between the latter tasks. This study examines how the distribution of the difficulty of tasks affects the improvement of test reliability of nested designs using generalizability theory. The research utilizes analog data to examine the influence of question group facet and item facet on test reliability in different difficulty distribution conditions. First, this paper analyzes the influence of the number of tasks and items on test reliability under the condition that there is no difference between the difficulty of task groups. Second, we analyze the influence of the number of tasks and items on test reliability when there is a difference between the difficulty of tasks. Results showed that the effective method to improve test reliability is to increase the number of tasks when there is a difference between the difficulty of tasks; and when there is a difference between the difficulty of tasks, the effective method to improve test reliability is to add items to each task. [29832]

56 Iowa Gambling Task in adolescents: Performance and Win-Stay/Lose-Shift strategy

Xiao-Xiao Liu, *Beijing Normal University, China*

The Iowa Gambling Task (IGT) is one of the most popular paradigms for decision making under uncertainty as well as affection decision making. Previous studies found critical performance developments in adolescence (Beitz, Salthouse, & Davis, 2014; Cauffman, et al., 2010), known for being impulsive and irrational. Compared with Reinforcement Learning models (e.g. Prospect Utility Learning model, PUL), Win-Stay/Lose-Shift (WSLS) strategy (Barto & Sutton, 1998) based on Heuristic model, sensitive only to the outcome of the previous choice, is recently approved a common used strategy in decision making. The present study focused on performance differences and strategy used between exploitation and exploration during adolescence. 347 middle school students aged from 12 to 15 were investigated. Personality measures including the Rosenberg Self-esteem Scale, State-Trait Anxiety

Inventory were administered before the gambling task. Cognitive modeling analysis shows that the WSLS model fit well to IGT data from our sample of healthy adolescents. No age differences were found for parameter γ_w (i.e., $P(\text{stay}|\text{win})$), while parameter γ_1 (i.e., $P(\text{shift}|\text{lose})$) decreased linearly with age. Consistent with previous research, adolescents with low self-esteem and high anxiety tend to have lower performance in IGT. [29806]

57 The application of scene simulation test in large-scale campus recruitment assessment
Ying Liu, *ATA, China*

The Scene Simulation Test (SST) is designed to evaluate candidates on situations that occur in various simulated workplaces. When taking SST, candidates are put in situations that are very similar to actual workplaces. They are required to solve multiple tasks through different items. The skill level, which is the most important factor to determine whether a candidate fits the job, is being tested during this process. In this study which was conducted in China in 2013, the number of examinees was over fifty thousand. The results of the study show that SST has better discrimination, better testing equality, higher construct validity, and higher face validity than traditional assessment, and predicts the actual performance of candidates in the future work more effectively than traditional assessment. For the first time in large-scale campus recruitment assessment, computer-based SST was used to assess basic work ability. The results suggest that SST can be a better tool than traditional assessment in a competitive selection process. SST should be the significant new trend in the assessment area. [29573]

58 Item difficulty modeling of English reading comprehension for Chinese students
Yanan Liu, *Beijing Normal University, China*
Fang Luo, *Beijing Normal University, China*

Based on Embretson's reading comprehension cognitive processes model, this study identified and quantified the sources of difficulty of reading comprehension items, and eventually modeled their item difficulties. Based on the model, the difficulty of the English reading comprehension items were predicted correctly. Embretson developed a cognitive processing model for reading comprehension, which consists of text representation and a response decision process. In Study 1, teaching and research staff in China simulated the process of picking an English text and making up questions with a verbal report. The stimulus features which they considered were defined from their report. Additionally, several Chinese students participated

in a reading comprehension test with a verbal report, from which the stimuli affecting the process of problem solving were defined. Researchers considered the stimuli received from the two different groups of subjects to establish the final item difficulty model. The model includes two types of features–basic and integrated features–influencing text representation and affecting the response decision. Data from eight reading comprehension texts (consisting of 40 items) taken by 2,000 Chinese students were analyzed in Study 2. The item parameter estimates were calibrated using testlet response theory. We coded the text features by text analysis and modeled item difficulty in terms of basic and integrated features. It turns out that the best model includes the integrated features and a few of the stimulus features related to a response decision process. Based on the model, researchers predicted the item difficulty of some new English reading comprehension items correctly. [29704]

59 Mixture cognitive diagnostic model with covariates for differential item functioning analyses
Yanlou Liu, *Beijing Normal University, China*
Tao Xin, *Beijing Normal University, China*
Wei Tian, *Beijing Normal University, China*
Qin Yi, *Beijing Normal University, China*
Fang Li, *Taishan University, China*

In Cognitive Diagnostic Models (CDMs), Differential Item Functioning (DIF) occurs when examinees from different groups or latent groups with the same attribute profiles have different probabilities of endorsing an item. The detection and analysis of DIF is critical in CDMs, since DIF is a threat to the validity and fairness of the test. Mixture models have been suggested as a useful tool to detect and analyse DIF. In contrast to DIF detection procedures based on manifest proxy groups, mixture models provide a latent approach to directly analyze DIF. We extended the CDM to more general situations and covariates are included in the model to predict each examinee's latent class membership. Here we use a log-linear cognitive diagnostic model as an example. We use the Expectation-Maximization (EM) algorithm to obtain maximum likelihood estimates of the model parameters. [29644]

60 Effects of distance and pattern on estimates of piecewise growth mixture model
Yuan Liu, *The Chinese University of Hong Kong, Hong Kong*
Fang Luo, *Beijing Normal University, China*
Hongyun Liu, *Beijing Normal University, China*

The Piecewise Growth Mixture Model (PGMM) builds on the Piecewise Growth Model (PGM) and the Growth Mixture Model (GMM), which is useful in detecting the non-continuous growing trend in a heterogeneous population. Factors affecting PGM and GMM will affect the estimates and performance of PGMM, such as the change of the slope, the distance between latent classes, and the sample size. Moreover, as a special and newly emerged issue in PGMM, even for models with the same distance, their different slopes can be combined to form different patterns. We used a simulation study to examine the effects of sample size, latent distance, and the pattern. The results showed that: (1) the distance between the latent classes (squared Mahalanobis distance, SMD) was a crucial factor that influenced the model selection and parameter estimation. A large distance would lead to consistent BIC and entropy when the right models were selected; while a small distance would not. (2) The pattern of the growth trajectory would affect model selection; specifically, non-parallel patterns of the trajectory would help model selection (higher entropy and higher total hit ratio) for medium distance (SMD = 3) and medium sample size (N = 200) conditions. (3) Parameter estimation was affected by the sample size and distance between latent classes. Parameter estimates would become more precise as the sample size and the distance increased. [29680]

61 Using multilevel multigroup item response theory models to better estimate nesting effects

Zhenqiu (Laura) Lu, *University of Georgia, USA*
Li Guan, *University of Georgia, USA*
Nathan T. Carter, *University of Georgia, USA*

Multilevel Item Response Theory (IRT) models hold great potential for use in educational settings to accurately model nested data. Recent advances allow for estimating the influence of nesting variables on individual-level scores (e.g., the influence of teachers on student scores), while taking account of differences in item parameters between groups (e.g., grade level). Here we apply the MultiLevel MultiGroup (MLMG; Yang, Monroe, & Cai, 2012) IRT model to estimate the influence of teachers on student scores. Results are compared to analysis without accounting for item parameter differences between grades. The MLMG model and multilevel IRT model (not accounting for group differences) were compared in the analysis of students nested within teachers, in one of three grades (6, 7, and 8). Students responded to 4 aspects of each learning-related question: Inquiry, Content, Everyday Language, and Academic Language. For grades 7 and 8, the MLMG model produced a larger Intraclass Correlation (ICC), suggesting that more student score variance was explained by the teacher, than the model that did not take into account item parameters differences between grades. The MLMG may lead to a better understanding of the importance of key variables in educational data by allowing researchers to faithfully model complex nesting. We will conduct and present a simulation that investigates whether accounting for group item parameter differences results in the estimation of more accurate ICCs and person parameter estimates in multilevel measurement models. [29877]

62 Application study of online multistage intelligent adaptive testing for cognitive diagnosis

Fen Luo, *Jiangxi Normal University, China*
Shuliang Ding, *Jiangxi Normal University, China*
Xiaoqing Wang, *Jiangxi Normal University, China*
Jianhua Xiong, *Jiangxi Normal University, China*

In long-term use of Computerized Adaptive Testing (CAT), people have discovered some defects of it. To offset its disadvantages, the Multi-Stage adaptive Test (MST) was proposed, which takes a group of items as a unit instead of a single item. Recently, On-the-fly MST (OMST) is addressed which combines CAT with MST. We combined Cognitive Diagnosis (CD) and OMST to build a new test design named CD-OMST, whose goal is to accurately estimate examinees' overall ability level and their Knowledge State (KS) simultaneously. CD assessment explores the examinees' cognitive profile, which is dependent on CD Model (CDM). According to whether the KS is included in the Item Response Function (IRF), we can divide CDMs into two types: the explicit CDM and the implicit CDM, for example, DINA is one of the explicit, AHM and GDD are two of the implicit. At present, the research on equating technologies on explicit CDM is insufficient. Most of implicit CDMs are built on Item Response Theory (IRT), IRT can provide sophisticated equating technologies to construct large item banks. The SHE and PWKL item selection indexes which were created on the basis of explicit CDM cannot be used in CD-OMST, so we built a new selection index CD-OMST which has nothing to do with any CDM, and the knowledge state is estimated by GDD. Monte Carlo simulations show that CD-OMST can gain a higher pattern match ratio and can estimate the examinee's ability level at the same time. [29669]

63 Method matters: An investigation of teacher rankings with different evaluation methods

Xin Luo, *Michigan State University, USA*
Huili Liu, *University of Maryland, College Park, USA*

After years of holding schools accountable for students' performance on large scale assessment, researchers and policymakers are trying to expand the usage of test scores to teacher evaluation. The current study tried to use two widely applied school evaluation models, the Student Growth Percentile (SGP) model and the Value Added Model (VAM) with covariate adjustment, to evaluate teachers. Within the SGP model, we applied 5 different evaluation methods, and within the VAM we applied 3 different methods. The same group of teachers was first ranked based on the aforementioned eight methods. Then we obtained the correlations among the eight sets of rankings to examine the agreement among the methods. In addition, as teacher evaluation results will have more profound impact on teachers who were ranked at the bottom, we took a closer look at teachers who were ranked among the lowest 10% by each method and compared the teacher lists suggested by each method. We found that teacher rankings were similar for some methods but different for others and the eight methods classified different teachers into the bottom 10%. We conclude that teacher evaluation can be influenced by different evaluation methods. Because different evaluation methods serve different evaluation purposes, when high-stake decisions were made, caution should be taken to make sure the method chosen serves the right purpose. [29880]

64 An introduction of Bayesian structural equation modeling in SPSS AMOS

Jingyao Mai, *Breadtrip, China*
Hongwei Yang, *China*

The study provides an overview of a Bayesian approach to Structural Equation Modeling (SEM) and illustrates the advantages of the approach over the traditional, frequentist practice. To that end, the study discusses typical complications in SEM (ordinal categorical and/or dichotomous data, missingness, etc.) and demonstrates how Bayesian SEM can be used to effectively handle such complications. SEM provides a framework for modeling relationships in multivariate data. The Bayesian approach to SEM is a relatively new methodological development rooted primarily in progress in computational algorithms using Markov chain Monte Carlo (MCMC) sampling. In Bayesian statistics, model parameters are treated as random quantities and hence are represented via probability distributions. A Bayesian model specification requires prior distributions for parameters. Then, given observed data, posterior distributions of parameters are obtained from which MCMC starts to generate random values for those parameters. Bayesian statistical inference is based on sampled parameter values from the posterior distributions. The combination of Bayesian statistics and SEM makes it possible to factor prior knowledge into the model to provide better insights into the substantive problems. The study uses SPSS AMOS to demonstrate important features of Bayesian SEM for an application from the Program for International Student Assessment (PISA). During the analysis, multiple structural models are built in a hierarchical manner and are evaluated for goodness-of-fit using Bayesian fit measures. Both non-informative and informative priors (from previous PISA data) are considered and compared in terms of parameter estimations. The analysis is performed in AMOS using its random walk Metropolis algorithm. [29720]

65 Effectiveness of two approaches to exploratory factor analysis of situational test data: ESEM and MBFA

Sunhong Min, *Sungkyunkwan University, South Korea*
Soonmook Lee, *Sungkyunkwan University, South Korea*

Item scores in situational test data include not only responses to a construct but also method effects generated by the situation, which creates a difficulty estimating the construct controlling for the method effects. For this reason, procedures for construct validation of situational tests have not yet been established. This study introduced two exploratory MTMM approaches that allow researchers to estimate measurement models while controlling method effects in different ways. One approach is the application of Exploratory Structural Equation Modeling (ESEM; Asparouhov & Muthén, 2009) and the other is Multiple Battery Factor Analysis (MBFA; Browne & Tateneni, 1997). To examine the performance of the two models, we conducted a Monte Carlo simulation where the study factors were the size of the method effects and the model type (ESEM, MBFA). Dependent variables were the ratio of convergence, ratio of proper solutions, and estimation accuracy and stability of factor solutions. For the ratio of convergence and the ratio of proper solutions, the results differed across conditions. When the method effects were small or medium, the ratio of convergence and the ratio of proper solutions of the two approaches appeared similar. However, when the method effects were big, MBFA performed much better. For accuracy and stability in estimation, MBFA showed better performance across all conditions. The results of the simulation study were supported in empirical data;

in data with small method effects, the two approaches produced similar solutions, but in data with big method effects, ESEM did not converge and MBFA produced better solutions. Suggestions are given following the results. [29801]

66 Different level factors of influence job satisfaction among elementary school teachers

Hong-Hua Mu, *Beijing Normal University, China*
Hongyun Liu, *Beijing Normal University, China*

This study examines the main factors of which influence job satisfaction among a sample of 13,398 teachers in 13,398 elementary schools in China. Hierarchical linear modeling was used to assess both school level factors as well as factors at the individual level of analysis in relation to job satisfaction after controlling the background characteristics at both the teacher level and school level. The results indicated that: (1) 24.7% of the total variance in elementary school teachers job satisfaction is between schools; (2) at the teacher level, after controlling for the effects of individual demographic variables (gender, age, and education attainment), the effects of teacher level factors such as wages, average daily workload, job vitality, and job absorption on job satisfaction were significant, and (3) at the school level, the effects of colleague relations, supervision, and evaluation on job satisfaction were significant. Based on the research, the paper further discussed advice on how to improve China's primary school teachers' job satisfaction, such as by enhancing teachers' job happiness, strengthening democratic decision-making, and accelerate the professional development etc. [29921]

67 Latent-shift model and its applications

Tomoya Okubo, *The National Center for University Entrance Examinations, Japan*
Kenske Okada, *Senshu University, Japan*

In this presentation, we discuss a Multidimensional Scaling (MDS) model that allows us to describe inconsistent human choice data. Describing human choice behaviour is an important topic in psychology and related fields. An approach to describe relations between stimuli and decision-makers is to analyse choice (preference) data using statistical models such as MDS models. Although many MDS models have been proposed to analyse human choices, they do not provide sufficient with information in real data analysis. A plausible reason for that is that MDS models cannot describe inconsistent choices, that they attribute to error. One approach for allowing

intransitive choice is to employ mixed-preference model (Okubo & Mayekawa, 2013). It is a MDS model that allows multiple preference orders within a subject, whereas classical MDS models for choice data assume only one preference order within the subject. Although the fundamental idea of modeling viewpoint shifts has been presented, some model generalizations can be achieved. In this presentation, we formulate a Bayesian solution for mixed-preference models. In addition, we show a post-processing method for the model in order to evaluate the posterior distribution of the parameters appropriately. [29797]

68 A new statistical method for analyzing children's longitudinal cognitive development based on a measurement invariance condition

Xiangzi Ouyang, *Beijing Normal University, China*
Wei Tian, *Beijing Normal University, China*
Tao Xin, *Beijing Normal University, China*

Cognition is not only the core of human intelligence but also an essential ability of children. In previous studies, longitudinal methods have frequently been used for analyzing children's cognitive development. However, few researchers have considered the measurement invariance problem in longitudinal studies. Thus, their results may be biased or even misleading. The purpose of our study is first, to use a newly developed MIRT method to analyze the longitudinal trend of preschool children's cognitive development; second, to analyze whether the instrument satisfied the measurement invariance condition or not; third to find items that did not satisfy measurement invariance, to help future revisions of the instrument. In this study, we chose 882 48-month-old children from the country and tested their cognitive development in the second year. After model comparison and considering the instrument construct, we choose a 2-tier item response model to analyze the instrument and to test the test-level measurement invariance and item-level measurement invariance. The result show that the instrument is partially invariant, which shows that the construct remains stable over age. On the other hand, items 2 and 3 did not satisfy the measurement invariance condition. Thus researchers can revise or delete these items to ensure the quality of the instrument. Moreover, the children's cognitive proficiency grows significantly from age 4 to 5. [29571]

69 A multilevel investigation of the factors influencing licensing deficiencies in South Carolina

Fan Pan, *University of South Carolina, USA*
Wenjia Wang, *University of South Carolina, USA*

To ensure the compliance of licensing regulations of child care facilities, licensing units make unannounced/announced visits. Licensing units also visit child care facilities (operating with or without license) when a complaint has been registered with the agency. During each visit, a licensing specialist collects information on any non-compliance of regulations and the licensing units refer to the non-compliance as "deficiencies". Any cited providers who fail to resolve deficiency would face a mandatory removal of their license. The literature on factors influencing the number of licensing deficiencies is sparse (Fiene, 2002). We start with this exploratory study to see the influencing factors of the number of licensing deficiencies. We can use the result to inform our work of providing on-site technical assistance so that to help the child care facilities resolve licensing deficiencies more efficiently. To explore the factors that may strongly influence the total number of deficiencies, we analyze data extracted from the latest review in the state licensing data from 2011 to 2014. We used Hierarchical Linear Modeling (HLM) as our statistical method, and we have two levels of factors: Level-1 for child care facility characteristics and Level-2 for county characteristics. We found that among the Level-1 factors, "Facility Type" significantly influence the total number of deficiencies. Among the Level-2 (county factors), "Total Number of Specialists", "Level of Poverty" and "Total Number of Regulated Child Care Facilities" in the county significantly influence the total deficiency number. [29827]

70 The development of the peer bully/victim problem during primary school: An application of latent transition analysis

Yiqin Pan, *Beijing Normal University, China*

The peer bully/victim problem changes considerably during the high grades in primary school. However, previous research about school-bullying problems lacks classification research focusing on both the bully and the victim and longitudinal studies focusing on transitions between the bully and victim. Therefore, this study applies Latent Class Analysis (LCA) and Latent Transition Analysis (LTA) to longitudinal data on 712 students in Grade 4, Grade 5 and Grade 6 to explore the categories of high-grade students in primary school and their transitions. The results of the LCA indicate that (1) students in Grade 4 to Grade 6 could be divided into four categories, including double-involved, bully, victim and non-involved; (2) non-involved students are the majority and double-involved the minority in all grades. In addition, the results of the LTA show that (1) the transition of students

is stable which means that going from Grade 4 to Grade 5 and going from Grade 5 to Grade 6 have similar transition probabilities; (2) compared to going from Grade 4 to Grade 5, when going from Grade 5 to Grade 6, students tend to transition to the categories with a lower degree of bully behavior; (3) the categories in Grade 4 will have a lasting impact on the categories in Grades 5 and 6. [29962]

71 3-level IRT models for educational data from the German National Educational Panel Study (NEPS) from secondary schools

Theresa Rohm, *Leibniz Institute for Educational Trajectories, Germany*
Claus H. Carstensen, *Leibniz Institute for Educational Trajectories, Germany*
Luise Fischer, *Leibniz Institute for Educational Trajectories, Germany*

A multilevel IRT model for reading competence data of grade 5 students, considering the stratified structure of the German school system will be investigated. Due to the hierarchical structure of students nested within classes and schools, and schools within different types of secondary schools which furthermore differ across federal states, the estimation of latent ability requires multilevel modeling. To estimate plausible values of latent competence and simultaneously account for missing values in background variables, a one-step estimation procedure was developed (Aßmann et al., 2014). Missing values are addressed through a MCMC based estimation strategy, with use of the Gibbs sampler (Fox & Glas, 2001). Predictors on the first level are the item-responses of students and given information on background variables. Second level units are determined through the school- and classroom-setting. Information about the observed school type is indicated on the third level. For this level the cluster assignment is fixed with cluster specific conditional means and conditional variances. The estimation accuracy of these fixed parameters will be evaluated. A further step will be to consider mixing components within clusters on the second level. Of interest is, if the effect of the stratified school system can be determined by the observable variables on the second and the first level. It will be tested if this additional information leads to a better model fit, i.e., reduced error in estimation of item parameters and reduced mean squared error on ability prediction, as expected from Adams, Wilson & Wu (1997). [29640]

72 Research on the relationship of work family conflict, self-efficacy and work engagement in middle school teachers

Yuan Rongman, *Beijing Jingyuan School, China*
Qi Jie, *Beijing Jingyuan School, China*

Teachers' work engagement not only influences teachers' job satisfaction and life satisfaction, but also impacts students' education. This article aims to investigate the relationship between teachers' work engagement, work family conflict, and self-efficacy. The article came to the following conclusions: the level of work engagement is high, WIF and FIW have a negative relationship with work engagement; teachers' self-efficacy is positively related to work engagement; and work family conflict not only impacts work engagement directly, but also impacts work engagement through teachers' self-efficacy. Thus we can reduce teachers' work conflict, and improve teachers' self-efficacy to improve teachers' work engagement, thereby providing a reference for middle school teachers' development. [29776]

73 Bayesian analysis of KBSID-III adaptive behavior data using a zero-inflated ordered probit model

Hye-Eun Seok, *Ewha Womans University, South Korea*
Su-Young Kim, *Ewha Womans University, South Korea*

Excessive zeros are frequently observed in response variables when we assess behavioral characteristics in development of children. For example, in the Korean Bailey Scales of Infant and Toddler Development-Third Edition (KBSID-III)-adaptive behavior test, a zero scale was recorded more often than other scales, such as 1, 2, or 3. A regular Ordered Probit (OP) model can be used for ordinal dependent variables. However, it is not appropriate to use an OP model with zero-inflated data. An OP model has a limitation when there are two distinct groups: genuine non-adaptive and potential adaptive groups. We applied a two-step Zero-Inflated Ordered Probit (ZIOP) model in a Bayesian framework to the KBSID-III-adaptive behavior data. In the first step, the non-adaptive group is separated using a probit model into a genuine non-adaptive group where adaptive behavior is observed due to undifferentiation and a potential adaptive group where adaptive behavior is not observed during the reported period. In the second step, an ordered probit model is applied to only the adaptive children group. A Bayesian estimation process for the ZIOP model is carried out using a Gibbs sampling algorithm. [29783]

74 Exploring the factor structure of the PISA teacher-student interaction scale based on Bayesian and traditional methods

Xuejun Shen, *Shanghai Academy of Educational Sciences, China*
Xiaowen Zhu, *Xi'an Jiaotong University, China*

In analyzing scales that address one single construct but include item parcels tapping similar aspects of the construct, commonly used models include the unidimensional model that forces a unidimensional structure to data, the correlated traits model where construct domain is broken apart into its correlated elements, and the second-order factor model where there is a second-order dimension explaining why primary dimensions are correlated. The bifactor model allows researchers to retain a goal of measuring a single common latent trait, but control for the variance that arises due to additional common factors (Reise, Moore, & Haviland, 2010; Reise, Scheines, Widaman, & Haviland, 2013). However, a bifactor latent structure appears to be not as commonly applied in the educational assessment field. The purpose of this study is to apply the unidimensional, correlated-traits, second-order factor and bifactor models to analyzing the Programme for International Student Assessment (PISA) Teacher-and-Student Interaction Scale data; estimate the model parameters and fit statistics using the Bayesian and traditional factor analysis methods; and compare the fit of the models to the data. The Teacher-Student Interaction Scale includes items that can be grouped into three aspects: teachers' instructional support, emotional support and classroom organization. Preliminary results show that by either estimation method, the bifactor model produces the best fit to the data among the four models compared. [29784]

75 Measurement invariance of the Behavior-Based Feeding Questionnaire across the U.S. and Taiwan

Yun Shi, *New York University, USA*

An estimated 15 million preterm infants are born every year globally and a high percentage of them exhibits feeding related issues. Thus, screening preterm infants' feeding abilities and exploring their feeding history are commonly included in routine follow-up visits. The use of feeding questionnaires with primary caregivers can help clinicians obtain feeding-related information from caregivers' perspectives, which provides an important basis for intervention. However, the perception of feeding problems is culturally influenced. In this study, I examine whether the Behavioral-Based Feeding Questionnaire (BBFQ) measures the same constructs, when administered to caregivers of either American or Taiwanese backgrounds. The data were collected from the United States

and Taiwan. A total of 177 observations were collected from Taiwan, and 198 responses for 77 infants were collected from the U.S. Exploratory Factor Analysis (EFA) was conducted using data obtained from the U.S. group, and the measurement invariance of the resulting EFA model was tested in the Taiwanese group. The results showed that the measurement invariance test supported the three-factor EFA resulting model structure, but no measurement invariance was found between two studied groups. In general, the factor loadings in the U.S. group are higher than those in the Taiwanese group. This supports the conclusions that cross-cultural differences may not reflect actual difference in clinical manifestation, but in caregiver's perceptions which is culturally influenced. Clinicians should be cautious when interpreting the BBFQ scores since their magnitude may be influenced by different cultures. [29908]

76 Number sense growth: A longitudinal study of children from third through fifth grade

Shu-Chuan Shih, *National Taichung University of Education, Taiwan*
Shu-Juan Lee, *National Taichung University of Education, Taiwan*

The main purposes of this research are applying the computerized number sense multiple assessment system for whole numbers and applying the best measurement model (higher-order item response theory, HO-IRT) developed in the previous study, implementing a longitudinal study from third through 5th grade, exploring the developmental trend of students' number sense ability after learning in school, and identifying potential growth trajectories for number sense ability of students. In this study, data for 3 time points were gathered by applying the computerized number sense multiple assessment system for whole numbers. Five assessments were administered to 334 students from third through 5th grade in Taichung. The valid sample size is 333. Real data from the assessments is analyzed by using HO-IRT, a Latent Growth Model (LGM) and a Growth Mixture Model (GMM) in order to explore the developmental trajectories of students' number sense abilities. The results are as follows: (1) In number sense and the four domain abilities, all growth trajectories of grade 3-5 reveal nonlinear de-accelerating growth curves. (2) Besides the ability of domain 1, there are significant individual differences in both the initial status and growth rate of number sense and the three domain abilities. (3) According to growth mixture models with different classes, there were no class differences in the patterns of growth trajectories in number sense and

the ability of domain 4. Two classes of growth trajectories are found in domain 1, 2 and 3. [29802]

77 Classical test theory analysis of students' perceptions of mathematics teachers in the Longitudinal Study of American youth (LSAY)

Mohammad Shoraka, *University of South Florida, USA*

Middle school students' perceptions of parents, peers, families, and teachers are widely viewed as being factors linked to student achievement. This study investigated the psychometric properties of questionnaire items used to measure students' perceptions of mathematics teachers in the Longitudinal Study of American Youth (LSAY) during middle school. The perceptions of students regarding their math teachers were gathered through 16 questions. The dataset was randomly split into two samples so that Exploratory Factor Analyses (EFA) could be conducted on one-half of the sample and Confirmatory Factor Analyses (CFA) could be conducted on the second half. EFA was deployed through Mplus and items were treated as categorical variables. Four factors were extracted under different methods of extraction within oblique rotations. Among these factors, two factors were common with two other studies: Instructional Strategies and Justice in Semmel (2007), and Instructional Strategies and Justice and Fairness in Sutcliff (2011). Cronbach alphas of Teacher Instructional Expectations and Teacher Fairness for this study were .74 and .62, respectively, in line with the two studies mentioned above for the instructional factor (.72 and .83) and for the fairness factor (.51 and .63). [29509]

78 The relation between dichotomous and polytomous reachability matrices and its applications

Ding Shuliang, *Jiangxi Normal University, China*
Luo Fen, *Jiangxi Normal University, China*
Wang Wenyi, *Jiangxi Normal University, China*
Xiong Jianhua, *Jiangxi Normal University, China*

In the design of the polytomous Q-matrix, a polytomous matrix (R_p) was proposed by Sun et al. (2013). They gave the statement that the matrix R_p can be used as the submatrix of the polytomous Q-matrix to guarantee bijective mapping from the set of Polytomous Ideal Response Patterns (PIRP) to the set of the Polytomous Knowledge States (PKS). But the statement was not proved in their study. In this paper, we give an expansion algorithm to make the translation from R_p to the dichotomous reachability matrix (R_2). Moreover, if the highest level of each attribute is known, we can give a compression algorithm,

which makes the translation from R_2 to R_p. Given the equivalent translations between R_p and R_2, one way to prove the above statement is to use the principle of design of the dichotomous Q-matrix proposed by Ding, Yang & Wang (2010). The PIRP is defined by the conditions that the relation between the attributes is non-compensatory and the scoring rule is to assign a grade based on the number of attributes mastered by KS (α) among the attributes being probed by item j. [29666]

79 The characteristics and influence factors of Chinese university students' values: A paired-comparison method

Xiaojuan Song, *Beijing Normal University, China*
Hongyun Liu, *Beijing Normal University, China*

In recent years, the education field has been concerned about the value orientation of university students. There is much theoretical research, as well as empirical rating scale research, focusing on the values of the Chinese. However, the responses of subjects to conventional rating scales are easily influenced by social desirability. Although the methods of paired-comparison and ranking can efficiently eliminate this negative impact, it has seldom been adopted in past research, especially in research about university students' values. To explore the characteristics and influence factors of Chinese university students' values, 548 university students were asked to rank 5 values, namely wealth, love, friendship, morality and ability, according to each value's importance to them. Then we recoded rankings into paired-comparison data, and used three cases of Thurstone models to analyze these data using Mplus. These models investigated value orientation in three different conditions: without individual differences (model 1), with individual differences (model 2), and with subject-covariates (model 3). The results indicated that model 3 with subject-covariate influencing the factor SES had the best model fit indices, and the following conclusions were obtained: (1) Overall, university students considered that morality was the most important, followed by ability, love, friendship, and wealth; (2) Individual differences did exist in university students' values; (3) SES could significantly influence values. Besides, gender and age had no significant effect on students' values. In this article, we also provided an alternative method to study value orientation. [29881]

80 Estimation of the initial value of the adaptive learning ability system

Mei Su, *Beijing Normal University, China*

In the adaptive learning system, a more accurate initial value helps to estimate the students' true ability level in a shorter period of time by using fewer testing items. This improves the adaptive learning system diagnostics, not only by saving diagnosis time and improving diagnostic efficiency, but also by helping to control exam exposure. Besides that, it can also improve personalized learning. There is currently little existing research on initial values for students across the different true ability levels in adaptive learning systems. Use of prior knowledge about the student can help to improve students' initial ability estimate. Thus, it is necessary to look for clues to design a method that will estimate students' initial capability uniquely in an adaptive learning system: such as how to mine and use student data in the system. As such, it is necessary to discuss how to efficiently use this prior information to help more accurately estimate students' initial ability. [29706]

81 Developing an instrument to assess cyber-bullying and cyber-victimization

Shaojing Sun, *Fudan University, China*
Xitao Fan, *China*

Measuring the construct of cyberbullying or cyber-victimization has been an intriguing issue in empirical social science research. Olweus (2012) reviewed published cyberbullying research, and suggested that researchers pay more attention to the definition and measurement of cyberbullying. Specifically, researchers need to make sure that they are studying the same phenomenon, and use similar measurement criteria (e.g., response alternatives, time frame for response). As Menesini & Nocentini (2009) noted, there was a severe lack of attention to measurement issues in the cyberbullying research. Some example measurement issues include the usage of global or unique questions for measurement, different understanding of this construct (i.e., cyberbullying) by different populations, and the difficulty in clearly defining the phenomenon because of the fast change of technologies. In the past, researchers have mainly adapted instruments targeting traditional bullying to measure cyberbullying. Validity and reliability of such adapted scales, however, have not been well established in the existing literature. Moreover, to date, most relevant instruments are developed or tested in a western cultural context. It remains unclear whether those instruments apply well to nonwestern cultures (e.g., the Chinese culture). The present study aims to develop an instrument to assess cyberbullying in the context of Chinese culture. Employing large-scale data collected online, the study will also provide insight

into the reliability and validity of the measurement. Use of a Chinese sample will contribute to the conceptual understanding of cyberbullying across cultures. [29446]

82 Formulation of GRM-based multilevel facets model

Xiao-Jian Sun, *Zhejiang Normal University, China*
Chun-Hua Kang, *Zhejiang Normal University, China*
Ping-Fei Zeng, *Zhejiang Normal University, China*

The purpose of this paper is to formulate an GRM-based model that can deal with the items whose step difficulties are monotonously increasing and can detect a variety of rater effects precisely and effectively when score CR items. This model, named Grade Response Multilevel Facets Model, contains two levels in this paper, the first level is an item response model, and the second level includes two regression models, one of which is related to variables that may affect the abilities of the person. The rest of the regression model is devoted to illustrating the rater effects. To examine the recovery of the parameters, two simulations were conducted. Simulation 1 uses a fixed-effects model such that only the person's abilities were modeled as random effects and both the effects of items and raters were fixed, meaning raters' attitude remains constant when rating the task-takers. In Simulation 2, both examinee ability and raters were set to random-effects such that raters can hold different standards to different test-takers. R software was adopted to generate examinee's responses for four items, and then the parameters of the model will be estimated by the SAS NLMIXED procedure. Results show that (1) there exists a little difference between the estimates and the true value in both conditions; (2) in both simulations, the model detects the rater effects precisely; and (3) the rater random-effect model is more suitable and stable than the fixed-effect model. [29550]

83 The determinants of training participation, a multilevel approach: Evidence from PIAAC round one countries

Teck Kiang Tan, *Institute for Adult Learning, Singapore*
Sheng Yee Zher, *Institute for Adult Learning, Singapore*

This paper uses logistic and Poisson multilevel models to examine the determinants of worker's participation in training programs using the Programme for the International Assessment of Adult Competencies (PIAAC) survey data covering 112,651 working adults in 24 countries. Five major forms of training are identified, namely taking up an open or distance learning, on the job training, attending seminars or workshops, taking up private lessons, and doing a formal education. Two measures of worker's participation in training are used: the number of training activities and whether respondents participated in at least one of these activities. We compare the results of complete pooling, fixed effect, random intercept, and random slope models, and point out the usefulness of random slope models in providing information about the differential effects on covariates for the countries. Logistic and Poisson multilevel models show consistent findings. The results show that socio-demographic and employment-related characteristics, and the skills used at work and in everyday life do affect the training participation of workers. Younger workers, those with higher education and more educated parents, those working in larger firms, and those working in information & communications, finance & insurance, business services, and community, social & personal services are more likely to participate in training than older workers, those with less education and less educated parents, and those working in small and medium-size firms. One crucial finding is that high use of literacy skills at work, literacy skills in everyday life, and numeracy skills in everyday life are strong determinants of training participation. [29570]

84 A factor mixture modeling approach for understanding the subtypes of posttraumatic stress disorder

Yun Tang, *Central China Normal University, China*
Li Wang, *Chinese Academy of Sciences, China*

The existence and identification of subtypes of Posttraumatic Stress Disorder (PTSD) has been debated (Dalenberg et al., 2012). With the advancement of statistics and technology, latent variable methods have been applied to provide evidence for PTSD subtyping. In the present study, the Chinese version of PTSD Checklist-Civilian (PCL-C, Wang et al., 2011) and a dissociative symptom checklist was administered to a sample of 362 Wenchuan earthquake survivors. The relationship between PTSD and dissociative symptoms was examined using Latent Class Analysis (LCA) and Factor Mixture Modeling (FMM). Preliminary LCA results preferred a two-class model. FMM was further conducted following the five-factor dimension of PTSD – intrusion, avoidance, emotional numbing, dysphoric arousal, and anxious arousal – plus dissociation. The best fitting 6-factor mixture model contained 2 classes, with invariant factor loadings, but varying factor covariance matrix and item thresholds between classes. This model suggested that differences between latent classes might come from the responses to single items, but not the underlying dimensions or mechanism of PTSD. Therefore, the subtypes of

PTSD may not differ qualitatively in their factor structures. We would also present the difference between the two classes in terms of coping strategies. [29727]

85 Bayesian estimation approaches for multilevel and multidimensional item response models

Jian Tao, *Northeast Normal University, China*
Jiwei Zhang, *Northeast Normal University, China*

For the multidimensional item response model with multiple levels, a Bayesian parameter estimation approach is developed by utilizing a data augmentation technique. A simulation study is conducted to show the parameter recovery of the new method. Also, a model selection criterion is provided for multilevel models. Finally, a real data example illustrates the implementation of the new method. [29595]

86 Equating subscores using total scaled scores as an anchor

Bing Tong, *Michigan State University, USA*

The great demands placed on subscores and scores on specific content areas/domains can be attributed to the potential remedial and instructional benefits they induce (Sinharay et al., 2011). Due to their small number of items, subscores have low reliability, but are highly related to the total/overall score. Equating subscores is one solution to make them comparable across different forms for subscore reporting. The working hypothesis in the present study is that with very few items used as an internal anchor to equate subscores, different levels of correlations between total score and subscores will lead to different levels of precision of subscore equating. The NEAT designs and Chained equipercentile equating method were applied in the current paper. A simulation study, in which two 60-item test forms with three contents and 5 common items for each content were generated in the framework of three dimensional item response theory model, was conducted to equate subscores using the Total Scaled Score (TSS) as an anchor with three different levels of correlations between the overall score and the subscores. It was found that the more the total score was correlated with the subscores, the more accurate was the equating of subscores. The performance of the equating was assessed using Conditional Standard Error of Equating (CSEE), weighted average of the CSEE (Avg SEE), Conditional Bias (CBias), weighted average bias (Bias) and the Root Mean Squared Deviation (RMSD). From the study, we may claim that, to a certain extent, using TSS as an anchor can be a valuable alternative to equating subscores. [29841]

87 A three-parameter speeded item response model: Estimation and application

Joyce Chang, *Academia Sinica, Taiwan*
Henghsiu Tsai, *Academia Sinica, Taiwan*
Ya-Hui Su, *National Chung Cheng University, Taiwan*
Edward M. H. Lin, *National Chiao Tung University, Taiwan*

When given time constraints, it is possible that examinees leave the harder items till later and are not able to finish answering every item in time. In this paper, this situation was modeled by incorporating a speeded-effect term into a three-parameter logistic item response model. Due to the complexity of the likelihood structure, a Bayesian estimation procedure with Markov chain Monte Carlo was presented and its performance on estimating the proposed model was demonstrated through simulations. For illustration purposes, the methodology was applied to physics examination data of the Department Required Test for college entrance in Taiwan. [29798]

88 Chinese people's cognition and attitude toward post-disaster psychological crisis intervention: Questionnaire development and validation

Peng Wang, *Shandong Normal University, China*
Junli Zhou, *Shandong Normal University, China*
Hong Dai, *Shandong Normal University, China*
Yiping Fan, *Shandong Normal University, China*
Fang Wang, *Shandong Normal University, China*

Our objective is to develop a Chinese people's cognition and attitude toward postdisaster psychological crisis intervention questionnaire and evaluate its psychometric properties. This study was comprised the following 3 stages: (1) Designing the questionnaire involved a qualitative study with interviews. (2) Administrating the pilot questionnaire and performing data analysis involved item analysis, Exploratory Factor Analysis (EFA), and internal reliability analysis. EFA was performed using the principal component analysis method, and the internal reliability was evaluated using Cronbach's alpha coefficient. (3) Implementing the formal questionnaire and Confirmatory Factor Analysis (CFA) entailed a study of 1,195 participants from Shandong and Sichuan Provinces in China. CFA was performed to assess the construct validity and fit indices according to the EFA. The questionnaire was divided into 2 parts. The scale of Chinese people's cognition of postdisaster psychological crisis intervention (Cognition scale) comprised 15 items and 3 subscales, namely theoretical cognition, professional cognition, and objective cognition. The scale exhibited adequate validity and internal consistency reliability (0.807).

The scale of Chinese people's attitude toward postdisaster psychological crisis intervention (Attitude scale) comprised 17 items and 3 subscales, namely obligatory attitude, constructive attitude, and positive attitude. The scale exhibited adequate validity and internal consistency reliability (0.845). The questionnaire is a reliable and valid measure for studying Chinese people's cognition and attitude toward postdisaster psychological crisis intervention. This instrument is easy to administer and may be used by postdisaster aid providers and researchers as an assessment tool for Chinese people who have experienced disasters. [29763]

89 Estimating classification accuracy and consistency indices for multidimensional latent ability

Wen-Yi Wang, *Jiangxi Normal University, China*
Li-Hong Song, *Jiangxi Normal University, China*
Shu-Liang Ding, *Jiangxi Normal University, China*
Fen Lu, *Jiangxi Normal University, China*

Although Multidimensional Item Response Theory (MIRT) has enjoyed tremendous growth, solutions for some problems remain unavailable. One problem is the estimation of classification accuracy and consistency indices under the MIRT. Yao (2012) has only focused on classification accuracy and consistency indices under MIRT based on total sum scores. Note that classifications made with the latent ability estimates are equally or more accurate than classifications made with total sum score (Lathrop & Cheng, 2013), at least for the one-parameter logistic model, the two-parameter logistic model, and graded response model. Building on two previously proposed classification accuracy and consistency indices for unidimensional item response models, this study aims to extend them to estimate classification consistency and accuracy of multidimensional latent ability from a single test administration. In a simulation study, the estimated classification accuracy and consistency were contrasted with the simulated classification accuracy and consistency estimates to determine their performances. The effectiveness of these indices was evaluated for different information matrices, numerical integration methods, test lengths (with each abilities of 10, 15, and 20 items, respectively), sample sizes (N = 2,000 and 5,000), and correlations between underlying ability dimensions (r = 0, .3, .5, and .8). [29675]

90 The neural correlates of biased responses: The method effects in the Rosenberg Self-Esteem Scale are associated with right amygdalar volume

Yinan Wang, *Beijing Normal University,China*
Feng Kong, *Beijing Normal University,China*
Lijie Huang, *Beijing Normal University,China*
Jia Liu, *Beijing Normal University,China*

Self-esteem is a widely studied construct in psychology that is typically measured by the Rosenberg Self-Esteem Scale (RSES). Nevertheless, a series of cross-sectional and longitudinal studies have suggested that a simple and widely-used unidimensional factor model do not provide an adequate explanation of RSES responses due to method effects. To identify the neural correlates of the method effect, we sought to determine whether and how method effects were associated with the RSES and investigate the neural basis underlying these effects. Method: 272 college students completed RSES and magnetic resonance imaging (MRI). Results: Behaviorally, method effects were linked to both positively and negatively worded items in the RSES. Neurally, right amygdalar volume positively correlated with the method effects associated with negatively worded items. Conclusions: The amygdala is involved in modulating negative affectivity and the behavioral inhibition system; therefore, the current study sheds light on the nature of the method effects that are related to self-report with a mix of positively and negatively worded items. [29500]

91 A comparison of different approaches to handling missing data in functional data analyses

Kimberly Ward, *Arizona State University, USA*
Hye Won Suk, *Arizona State University, USA*

The emergence of sophisticated measuring devices such as functional neuroimaging, electroencephalogram, and motion capture equipment has enabled researchers to collect densely measured data that can be considered to arise from underlying smooth curves. Functional data analysis aims to analyze curves and surfaces varying over a continuum, and thus provides a useful framework to analyze such types of data. A classical assumption of functional data analysis is data are sampled at all measurement occasions (i.e., no missing data). This assumption is sometimes impractical as researchers encounter missing data for various reasons. Missing data can greatly bias the results obtained from analyses, in turn, leading to inaccurate conclusions. However, little work has been done developing and comparing methods in functional data analysis to handle missing data. This study aims to explore methods for handling missing data in functional data analysis. We draw on techniques proposed for both functional data analyses and for standard multivariate procedures including imputation via multilevel multiple imputation, random-effects models with basis

function expansions, and smoothing with basis function expansions. We examine various factors such as sample size, number of measurement occasions, degree of missingness, and type of missingness to evaluate the robustness of these methods across various conditions and discuss implications. [29750]

92 The theory and practice of personality development measurement

Weidong Wang, *Guang'anmen Hospital, China*
Jinhua Zhang, *Guang'anmen Hospital, China*
Xueyu Lü, *Guang'anmen Hospital, China*
Fan Feng, *Guang'anmen Hospital, China*

Based on abnormal personality development theory (in order to determine the architecture of memory-tracing development level) we developed the memory-tracing development level questionnaire from the perspective of personality theory of Chinese medical psychology and investigated different subject populations to understand the differences. Through a literature review of traditional Chinese medicine and psychology, we use literature analysis, qualitative research and brain storms to extract the theoretical basis, so as to build the architecture of the questionnaire and enact items. We also assessed the reliability and validity of the questionnaire by means of exploratory factor analysis and confirmatory factor analysis, t-tests, etc. Results are as follows: (1) The developed "Memory-tracing personality development level questionnaire" may be divided into sections of childhood, adolescence, and adulthood, where the final questionnaire is composed of 9 subscales, 40 dimensions and 275 items; (2) The reliability of the whole questionnaire is 0.967, and the reliabilities of the 9 subscales are between 0.779 and 0.981. The RMSEAs of all subscales are less than 1, with NNFI and CFI nearly 0.90, which indicate good quality of the questionnaire. Conclusion: (1) Chinese medical thoughts, holism, dialectical view and eternal movement, have played a guidance role for the memory-tracing psychological development level questionnaire, and the results of reliability and validity test support this conclusion. (2) The "Memory-tracing personality development level questionnaire" was demonstrated to possess good internal consistency, construct validity, criterion-related validity and congruent validity. [29854]

93 Is the recovery of latent class membership highly consistent by applying two types of constraints in mixture Rasch modeling?

Yi-Jhen Wu, *University of Bamberg, Germany*

Mixture IRT modeling is a promising approach, which is a combination of Item Response Theory (IRT) and Latent Class Analysis (LCA). It can overcome disadvantages of IRT and LCA, but also keep advantages of IRT and LCA. Therefore, it has been used extensively in education and psychology fields. In the literature, most researchers use conventional constraints, which are (i) to set all classes to be the same normal distribution with equal means, or (ii) to set the sum of item difficulty to be 0 for the Rasch mixture model. However, these constraints are not sufficient to make latent classes comparable when a test does not have anchor items (i.e. class-invariant item). Therefore, the aim of the simulation study is to investigate if recovery of class membership is consistent between the conventional constraint and the class-invariant items constraint which allows comparable latent classes. Results show that the recovery of class membership is highly consistent under the conventional constraint and class-invariant items constraint. [29535]

94 The multilevel DINA model

Shibei Xiang, *Beijing Normal University, China*
Wei Tian, *Beijing Normal University, China*
Yi-Bo Wang, *Beijing Normal University, China*
Panpan Han, *Beijing Normal University, China*
Tao Xin, *Beijing Normal University, China*

Cognitive Diagnosis Models (CDMs) attempt to find the latent traits or attributes that examinees must possess in order to answer the items correctly. The DINA (deterministic input, noisy, "and" gate) model has been widely used in cognitive diagnosis tests and in the process of test development. However, many datasets in education are in a hierarchical form where students are nested within schools, schools within school systems, and so on. This study aimed to extend the DINA model by considering the hierarchical form to allow examinees to take the dependency between observations based on the same cluster into account by introducing one or more random effects. A logistic version of the DINA model can easily be fit with hierarchical form for latent class analysis using non-linear mixed model framework for item response theory. It will be shown that the parameters of the new class of DINA model can be estimated in a Bayesian framework using Markov chain Monte Carlo methods with the computer software WinBUGS. Examples using simulated data are given to demonstrate that the model parameters and attribute-mastery profiles can be recovered relatively accurately using the new class of DINA models. Finally, mathematics data from PISA will be used as an empirical example to demonstrate applications and implications of the new class of models. [29821]

95 Bayesian structural equation modeling approach to confirming testlet effects: Is a bifactor structure necessary?

Xin Xin, *University of North Texas, USA*

Bifactor models have a long history yet have been controversial (Reise, 2012). In bifactor models, each indicator loads on a general factor and one of the group factors, and thus the main effect and secondary effects are evaluated separately. Despite the useful applications in dimensionality and subscore scaling, bifactor models are susceptible to computational problems. Bifactor models constrain all factors to be orthogonal for identification, but fixed-zeros might not be helpful, and may result in empirical underidentification and nonconvergence (Kenny, 1979; Wothke, 1984; Kenny & Kashy, 1992). Even with successful estimation, secondary effects are confirmed by a series of model comparisons based on model fit indices (DeMars, 2013), yet the rules of thumb are also controversial (cf. Kline, 2011). A specially designed Bayesian approach may solve this problem. The Bayesian Structural Equation Modeling (BSEM) approach (Muthén & Asparouhov, 2012) specifies cross-loadings and residual variance correlations as approximate zero priors instead of exact zeros. Unlike the traditional Maximum Likelihood (ML) approach, the Bayesian approach identifies models by small variance priors, therefore underidentified models in ML can be estimated in BSEM. Additionally, Bayesian methods assess model fit via posterior predictive checking (PPC; Gelman, Meng, Stern, & Rubin, 1996) and the multivariate potential scale reduction factor (MPSRF; Brooks & Gelman, 1998), therefore bypassing controversies of rule-of-thumb model fit indices. The present paper used TIMSS student survey data from multiple countries to demonstrate cases when ML produced Heywood cases and BSEM did not, and the ones in which only BSEM can assess the misspecification and the necessity of a bifactor structure. [29728]

96 Modeling online social networking behaviors: A multidimensional mixture IRT approach

Kuan Xing, *University of Illinois at Chicago, USA*
Yoon Soo Park, *University of Illinois at Chicago, USA*

Traditional Item Response Theory (IRT) models assume that person ability is a unidimensional latent trait. However, theoretical and practical applications in measurement require more advanced IRT models to handle multidimensional latent traits and potential latent subgroups that differ in response patterns. Prior studies (e.g., Kelderman & Rijkes, 1994) have proposed multidimensional IRT models, while others (e.g., Cohen & Bolt, 2005; Li, Cohen, Kim, & Cho, 2009) have separately proposed mixture extensions to unidimensional IRT models. While multidimensional extensions to mixture IRT models are under development, additional studies that are motivated from real-world measurement characteristics may improve the practicality of such models. This study proposes a multidimensional mixture IRT (MMIRT) model, which allows parameterizing multidimensional latent traits and identifying different latent subgroups of respondents. Real-world data from an online survey (51 items) measuring respondents' use of online social network sites were used to examine the proposed MMIRT. Empirical analysis showed that a 3 dimensional model with 2 item parameters (difficulty and slope) and 3 latent subgroups of respondents best fit the data; this model was better fitting than unidimensional IRT models (no class) as well as the unidimensional mixture IRT models (3 class). Furthermore, simulation studies were conducted to examine the recovery of parameters. Overall, this study shows that the MMIRT can be a useful model to estimate multidimensional traits while identifying latent subgroups of respondents. Applications of this model may help researchers and educators better understand how different groups of respondents or students behave in different dimensions. [29737]

97 The preliminary construction of a competency model of R&D employees in enterprises

Xinli Zhu, *Shandong Normal University, China*
Miao Guo, *Shandong Normal University, China*
Fengqiang Gao, *Shandong Normal University, China*
Peng Wang, *Shandong Normal University, China*
Jie Hu, *Shandong Normal University, China*

Our objective is to construct a competency model of R&D employees in enterprises. First, the superior group ($n = 11$) and ordinary group ($n = 11$) who were R&D employees in enterprises in Ji Nan were selected based on certain standards to conduct a Behavior Event Interview (BEI), which contained three successful and failed experiences respectively. Then we analyzed and encoded the interview recording. Result: (1) The index of category agreement, intercoder reliability and Kendall's coefficient of concordance as rater consistency indicators were all within an acceptable range, indicating a rater consistence for two groups. (2) The correlation between the frequency of competency, mean grade, supreme grade and the length of interview was not significant, which indicated that the differences of competency in frequency, mean grade and supreme grade were not caused by the length. (3) The employees' core competency were:

achievement orientation, self-confidence, optimism, monitoring capability, overall conception, capability against stress, spirit of service, spirit of sacrifice, building relationships, and flexibility. The basic competencies were: insight, sense of belonging to enterprise, sense of responsibility, humble attitude, persistence, innovation ability, research capacity, communication capability, logical skills, ability to influence, information aggregation ability, initiatives, honesty and integrity, communicative competence, organization awareness, leadership, cooperation. In conclusion, the competency model has high validity and reliability and can predict employees' performance, which differentiates excellent staff from others. [29822]

98 Response to tier 2 interventions: What makes a valid classification?

Qinfang Xu, *Beijing Normal University, China*

This study was conducted to examine whether English word reading, relevant cognitive skills, or their combination would be a better criterion in defining responsive status. Readers at risk (n=72) and typical readers (n=84) were screened from 668 fourth graders of four primary schools in Beijing by their performances on tests of word recognition, school English test scores, and teacher's nomination. The readers at risk received a 600 minute intensive training in phonological awareness and letter sound knowledge. A series of English word reading tests and relevant cognitive skills (phonological awareness, letter sound knowledge, rapid naming of letters, phonological memory) were individually administered 3 times: before, right after, and 1 year after the intervention. While significant differences were revealed on word reading and cognitive skills between typical readers and readers at risk one year later (no matter what their responsive status to intervention was), the differences between responders and non-responders were significant only when the responsive status was classified by word reading. Ordinal logistic regression models showed that word reading and cognitive skills predicted students' responsive status to intervention with a hit rate of 80.6% (sensitivity: 91.2%, specificity: 71.1%). However, no cognitive skill explained the unique variance in reading at re-test after accounting for pre-test word reading level. Word reading measures may be better criteria to define responsive status to Tier 2 interventions for poor English readers. And it would be necessary to apply IRT to understand the individual differences in learning to read English for Chinese children. [29767]

99 A component-based approach to find outcome-related clusters

Michio Yamamoto, *Kyoto University, Japan*

Atsushi Kawaguchi, *Kyoto University, Japan*

Cluster analysis, in which objects with various features are partitioned into several unpredetermined homogeneous groups (clusters), is one of the major exploratory multivariate analyses. Practically, the cluster analysis is followed by considering the meaning of the obtained cluster. Then, researchers hope that the cluster is related to some outcomes that they are interested in. For example, in the study of medical genetics, the cluster structure given by genome data is expected to be related to the degree of clinical disease. Also, in marketing research, the clusters are expected to relate to consumer behavior. However, it is often difficult to find the meaningful clusters since cluster analysis is essentially an unsupervised classification technique, that is, the cluster structure is estimated without referring to the interested outcomes. Thus, in this study, we propose a new clustering technique that can provide a partitioning related to some outcome variables. The clusters from the method would be easy to understand since the result can be interpreted from the outcome information. In the method, the aim is attained by a simple transformation of exploratory variables using outcome variables. Also, the proposed method is a component-based technique, where component scores that express the cluster structure clearly are estimated. Furthermore, we study a modification of the proposed method to deal with moderately high-dimensional data. [29589]

100 The mediational roles between state narcissism and aggression

Chenchen Yang, *Beijing Normal University, China*
Caina Li, *Shaanxi Normal University, China*
Yufang Bian, *Beijing Normal University, China*

In most prior research, narcissism has been consistently found to promote aggression under provocation. However, most research has focused only on dispositional narcissism without considering situational factors that may increase narcissism temporarily. The present studies explored the possibility that experimentally induced increases in narcissism would foster aggressive responding to criticism by increasing threat perception in a laboratory setting among Chinese college students. 162 college students were randomly assigned to conditions of a 2 (narcissism activating story vs. neutral control story) by 2 (threat vs. praise) between-subjects design. Structural equation modeling and bootstrapping were utilized to test the hypothesized mediating model. Results showed that there is a significant difference in the level of state narcissism induced by different priming stories

among participants; the state narcissism activating group expresses more aggression under the condition of threat (vs. praise); the effect of state narcissism on aggression is mediated by threat perception, anger and hostile attribution biases; threat perception not only directly predict aggression, but also lead to aggression via feelings of anger and hostile attribution biases. These results suggest that emotional and cognitive factors would produce aggressive behaviors when narcissists are confronted with negative evaluations. [29819]

101 Investigation of the psychometric properties of standard setting methods using multivariate generalizability theory

Heewon Yang, *Yonsei University, South Korea*
Yeonbok Park, *Yonsei University, South Korea*
Guemin Lee, *Yonsei University, South Korea*

For criterion-referenced evaluation, the cut-off score should separate a group of examinees into mastery learner and non-mastery learners. Various standard setting methods have been devised for setting the cut-off score, and the most frequently used methods are tthe Angoff and Bookmark methods. However, there have been relatively few studies in the literature on the reliability of cut-off scores. Cut-off scores can be determined differently by applying different kinds of standard setting methods. Consequently, examinees with the same score can be classified into different categories. The information on the reliability of the cut-off score could be one important consideration when choosing an appropriate standard setting method. The primary purpose of this study is to investigate the psychometric properties of both the Angoff and Bookmark standard setting methods. The specific research objectives are as follows: (1) To compare the Standard Errors of Measurement (SEM) of cut-off scores from the Angoff and Bookmark methods under a generalizability theory framework. (2) To compare the disattenuated correlations for universe scores between levels from Angoff and Bookmark methods under a multivariate generalizability theory framework. (3) To investigate optimal conditions for measurement following influence of variance component from the Angoff and Bookmark methods. (4) To compare the difference from standard error of measurement of cut-off scores and disattenuated correlations between levels among the subject. [29729]

102 Utilizing non-ignorable missing data information in item response theory

Ji Seung Yang, *University of Maryland, College Park, USA*
Daniel Lee, *University of Maryland, College Park, USA*

Any missing data is a nuisance in data analysis because the missing values introduce loss of information that could otherwise be used for better estimates of unknown parameters. The purpose of this study is to investigate ways to make use of this missing information to obtain parameter estimates within the item response theory framework. A simulation study is conducted to explore three different methods to incorporate missing data information under the assumptions of missing not at random (NMAR). Item response data with missing responses are generated using a mechanism that creates missing values based on three levels of correlations between the trait of interest and the propensity to have a missing response. First, a unidimensional nominal model with an extra category added to measure the propensity of missing is explored. Second, a two-dimensional model in which the additional dimension represents the propensity to have missing data is explored. Third, a multiple imputation approach that utilizes multiple sets of plausible item responses is explored. The imputation model uses observed mean scores and proportion of missing as predictors. All three methods are compared to a baseline unidimensional model where information about the missing data is not incorporated in the analysis. Results show that using information about the missing data is beneficial in reducing bias and standard errors of item and person parameters. These methods are also tried on a subset of empirical data (Trends in International Mathematics and Science Study, TIMSS) to demonstrate a valid reason to incorporate missing information. [29755]

103 Will self-image distortion in narcissistic people distort the validity of self-reported empathy measures?

Xi Yang, *New York University, USA*
Yuxiao Zhang, *New York University, USA*

With rising awareness of the global narcissism epidemic, the potential cure, empathy, draws much attention. Although empathy boasts a large number of self-report measures based on its complicated construct, ironically, their validity can be threatened by reporting bias due to the egocentric self-image in narcissistic populations. An extensive search was conducted for peer-reviewed literature from English- speaking countries with keywords 'empathy', 'scale', 'questionnaire', 'inventory' and 'measure' in the PsycINFO database from 1940 to 2015. The ten most frequent measures were evaluated on their comprehensiveness and median number of years of use. The Interpersonal Reactivity Index (IRI) by Davis and the Empathy Quotient (EQ) by Baron-Cohen and Wheelwright were selected. Eleven narcissism studies involving these

empathy measures were further reviewed. For IRI, the overall score was rarely used, with one of the two studies revealing a significant negative correlation with maladaptive narcissism. The most-used subscale, Empathic Concern, displayed non-significant negative correlation in three out of the six studies and Perspective Taking in two out of three. Similarly for EQ, two from the four studies were short of significance. The current popular empathy measures displayed inconsistent validity within narcissistic populations. The discriminant defect of the measures seems worse with sub-clinical narcissistic test-takers. More scenario based empathy tools and behavioral measures may be used to adjust the reporting bias of this participant group. The development of narcissism-adjusted empathy measures is promising. [29935]

104 Autoregressive latent growth modeling: A Bayesian approach

Yuzhu (June) Yang, *University of California, Merced, USA*
Sarah Depaoli, *University of California, Merced, USA*

The latent growth model and the autoregressive model are frequently used to answer different questions in longitudinal data analysis. Latent growth models, including Growth Curve Models (GCM) and Growth Mixture Models (GMM), allow one to examine individual variation in growth and to identify potential factors that explain the variation. In contrast, AutoRegressive (AR) models examine the time adjacent relations; it is assumed that an observed variable is a function of its immediately preceding value. The Autoregressive Latent Trajectory (ALT) model (Bollen & Curran, 2004) was developed to synthesize the latent growth (curve) model and the autoregressive model. However, the current specification of the ALT model does not allow for latent classes or non-linear growth trajectories. In this study, we extended the ALT model and examined the estimation accuracy through simulation of a series of linear and non-linear GCMs and GMMs with autoregression specified between the repeated measures data. We examined the impact of different forms of Bayesian priors to assess whether the estimation accuracy of the autoregression components and other parameters would be improved through the use of the Bayesian framework. The main findings indicated that the Bayesian estimation conditions performed better than the frequentist approach in terms of model convergence and estimation bias. Although the Bayesian framework yielded more accurate results, findings suggested that a sensitivity analysis of priors is necessary for understanding their impact on final model estimates. Finally,

we provide specific recommendations for the prior specifications for the current model. [29786]

105 Mixture IRT model with covariates for differential item functioning analyses

Qin Yi, *Beijing Normal University, China*
Yanlou Liu, *Beijing Normal University, China*
Tao Xin, *Beijing Normal University, China*
Tao Yang, *Beijing Normal University, China*
Hua-Xia Li, *Beijing Normal University, China*

If the experiences or cognitive processes leading to DIF are not perfectly correlated with the manifest groups, mixture IRT models can be more informative regarding latent groups. It has been suggested as a useful approach to detect and analyze the underlying DIF. Covariates can improve parameter estimation and help predict each examinee's latent group. However, the effect of covariates has never been systematically explored. The purpose of this study was to assess the performance and accuracy of a mixture IRT model with examinee covariates. We focused on model estimation and explored the effects of including covariates for latent class membership. In particular, we used the Expectation-Maximization (EM) algorithm to obtain item parameter estimates for each group and Wald tests to detect and analyze the underlying DIF accounting for the effect size. In a simulation, we compared 2 and 3 latent groups with different covariates, finding that the inclusion of covariates had a positive effect on the accuracy of the recovery of the underlying structure. It was shown that the mixture IRT model with covariates was not only a comprehensive approach for detecting different kinds of DIF, but also a feasible way of exploring possible explanations of DIF in the context of IRT. [29665]

106 Influence of elementary school student and teacher gender on achievement

Yi Tian, *Beijing Academy of Educational Sciences, China*
Yunfeng Wang, *Capital Normal University, China*
Yi Hao, *Beijing Academy of Educational Sciences, China*
Meijuan Li, *Beijing Academy of Educational Sciences, China*
Yongmei Zhang, *Beijing Academy of Educational Sciences, China*

Gender differences refer to the differences between males and females in capability, temperament, attitude, interest, behavior and other areas. It is one of the key issues for research in psychology and education worldwide. This study was conducted to explore the influence of the elementary school student and teacher gender on students'

achievement and to improve current education policies. Based on data from Beijing Assessment of Educational Quality (BAEQ), this study used cross-classified multi-level modeling. Participants included 99162 Grade 5 students and 3181 Grade 5 Chinese teachers, and 19766 Grade 5 students and 569 Mathematics teachers. The hierarchical linear modeling software package MLwin2.10 was used for data analysis. The following models were mainly considered: Model 1 (null model); Model 2 (add student gender variable and teacher gender variable to null model); Model 3 (add the control variables to model 2); Model 4 (add the interaction variables (student gender $\times$ teacher gender) to model 3). The results showed that student gender and teacher gender have a significant influence on student achievement in Chinese and Mathematics. Girls' achievement is higher than and boys', the students taught by female teachers have better achievement than those by male teachers. Student gender and teacher gender interact for student achievement in most components of Chinese, yet no interaction exists for the influence on achievement in any component of Mathematics. The results provide an empirical basis for developing teacher education policies and implementing gender differentiated teaching in basic education. [29756]

107 Validation of the Chinese version of the Immersed Tendency Questionnaire

Tian Yu, *Shandong Normal University, China*
Yulong Bian, *Shandong Normal University, China*
Xiaoyue Li, *Shandong Normal University, China*
Xinli Zhu, *Shandong Normal University, China*
Fang Wang, *Shandong Normal University, China*

The purpose of this study was to determine the factor structure of the Immersed Tendency Questionnaire (ITQ) in a Chinese cultural setting. The 14 items comprising the ITQ originally derived from Witmer and Michael's immersed tendency questionnaire (1998). Data from 350 university students were subjected to Exploratory Factor Analysis (EFA) and 431 university students were subjected to Confirmatory Factor Analysis (CFA). CFA confirmed that the 4-factor structure was better than the 3-factor structure. EFA explained 50.65% of the variance through a 19-item, 4-factor structure (game immersed, state involvement, focus, and emotional involvement). For the total score, Cronbach's alpha coefficients were 0.512-0.789. ITQ had a satisfactory structure validity (χ^2/df $=2.736$, NNFI $=2.736$, IFI$=0.93$, CFI$=0.93$, RMSEA$=0.064$) and criterion validity (r$=0.575$, p<0.001). Although a 3-factor solution was most defensible, with further refinement and additional items, the 4-factor had

a significantly better fit, and could be used in related fields of research. [29690]

108 The impact of ICT use and complexity thinking tendency on computer-based problem solving competence

Jianlin Yuan, *Beijing Normal University, China*
Hongyun Liu, *Beijing Normal University, China*
Fang Luo, *Beijing Normal University, China*

Student ICT use and complexity thinking have become a topic of interest in educational research (Eickelmann, 2011). Besides learning how to use ICT and the characters of complexity thinking, it is important to understand how ICT use and complexity thinking affect computer-based problem solving; as yet, it is not clear. Based on data from Singapore, Korea, Japan, and Shanghai-China in the Programme for International Student Assessment (PISA) 2012, the current study examined the impact of student's ICT use, complexity thinking tendency, and the richness of school ICT resources on computer-based problem solving competence. Two-level structural equation modeling was used and the results indicated that, at the student level, complexity thinking tendency positively predicted computer-based problem solving competence, whereas the frequency of ICT use negatively predicted computer-based problem solving competence. At the school level, richness of school ICT resources does not significantly predict computer-based problem solving competence. At the same time, we tested the difference between the four countries and regions using a multiple group comparison model and provide an in-depth discussion of the difference. We conclude that, for improving students' computer-based problem solving competence, strengthening the training of complexity thinking is more important than only increasing the frequency of ICT use. [29719]

109 Multiple matrix-sampling and its use in TIMSS 2015

Ma Yue, *Beijing Normal University, China*

Matrix sampling, sometimes referred to as a split-questionnaire, is a sampling design that involves dividing a questionnaire into subsets of questions, possibly overlapping, and then administering these subsets to different subsamples of an initial sample. This design reduces the data collection costs and addresses concerns related to response burden and data quality, but also reduces the number of sample units that are asked each question. This paper first provides an overview of the current research on multiple matrix sampling, reviewing its origins,

highlighting the fields besides education in which it has received the most application and discussing how it has been applied to problems in surveys. Second, I will refer to the specific survey process of matrix sampling and its restrictions and considerations by taking TIMSS 2015 as an example. Finally, the paper will conclude with practical experiences to enhance the overall large-scale assessment level in China and an identification of future work. [29714]

110 Study of the effects of types of motives to learn on student engagement: Based on data from the Chinese College Student Survey (2013)

Yuheng Huang, *Tsinghua University, China*

Intrinsic and extrinsic types of motivation are two important theoretical constructs in educational studies. They were defined as sources of motivation to learn. But it is argued that these two types of motives have different effects on learning practice. In this study, data from the Chinese College Student Survey (2013), which includes 64 four-year colleges in mainland China, were used to analyze the influence of different types of motives of higher education students. The results indicated that the most important source of motivation of Chinese college students is pressure of employment and learning career, which was reported by 43.85% of the students. Other sources of motivation were challenge/self-promoting (24.21%), interest in exploring knowledge (14.09%), quality of curriculum (6.86%), getting higher scores (5.62%) and pressure of parents/teachers (5.36%). It also revealed that students who are motivated by intrinsic motivation engaged more in college learning activities when compared with students who were mainly motivated by extrinsic sources of motivation. Significant t-tests suggested that the former type of student has better performance in the five benchmarks (Level of Academic Challenge, Active & Collaborative Learning, Student-Faculty Interaction, Enriching Educational Experiences and Supportive Campus Environment) of the survey. [29830]

111 The effect of home activities on math achievement for K-5th grade children: A latent growth curve modeling study

Duan Zhang, *University of Denver, USA*
Courtney Tobiassen, *University of Denver, USA*

Parenting practice has been proved to substantially affect children's academic success. Despite of the tremendous amount of attention math has received in educational research, few literature has explored how home activities would influence children's math achievement. The

current study aimed to understand the impact of home activities on math growth trajectories from kindergarten through fifth grade with a nationally representative longitudinal sample. The secondary data used for analysis came from the Early Childhood Longitudinal Study – Kindergarten Cohort (ECLS-K) program sponsored by the Institute of Education Sciences. The sample size was 9,752. Data were available at kindergarten, 1st, 3rd, and 5th grades. Exploratory factor analysis followed by Rasch analysis was used to construct a home activity composite based on 9 items from parent interviews. Then the home activity composite was modeled as a predictor to affect the intercept and slope estimated by the growth trajectories of math achievement using Latent Growth Curve Modeling (LGCM). A group of competing unconditional LGCM models were explored and the two factor linear growth model was found to fit the data best. The home activity composite in kindergarten was a significant predictor for both the initial status and linear growth rate of math achievement, indicating that parents can have greater impact on math achievement growth by doing activities with their children in earlier years. Methodological concerns of using LCGM and Rasch models for large scale national data were also discussed as part of the study emphasis. [29952]

112 What decides when marriage ends in the US: An analysis of marriage survival through Bayesian censored regression modeling

Qun Zhang, *University of Kentucky, USA*
Hongwei Yang, *University of Kentucky, USA*

This paper aims to ascertain the factors that contribute to divorce in the US. To that end, the paper conducts a survival analysis where the event is defined as *ended up divorcing*, and censored regression modeling is used to predict the amount of time to that event using couples' marital conflict resolving strategies, perceived fairness in domestic tasks, and time spent together. The regression analysis is estimated under a Bayesian method (with Markov Chain Monte Carlo, or MCMC, sampling) to take into account prior information from existing research in the field of family sciences. The reason for the choice of censored regression is that the dependent variable is right-censored: Some couples still had not divorced when we last observed them at the end of the sixth year. During the analysis, multiple models are built in a hierarchical manner and evaluated for goodness-of-fit using Bayesian fit measures (posterior predictive *p* value, etc.). Bayesian statistical inference is conducted on the final model to describe the relationship between the duration to the event

(i.e., divorce) and the independent variables. The dataset is obtained primarily from the National Survey of Families and Households (NSFH), and the analysis is performed under the Bayesian SEM platform of SPSS AMOS using its random walk Metropolis algorithm. [29967]

113 Comparison of Bayesian model selection methods for multilevel IRT models
Xue Zhang, *Northeast Normal University, China*

This study presents new findings on the utility of DIC as a model selection index for dichotomous Multilevel Item Response Theory (MLIRT) models. The MLIRT model consists of an IRT model, an individual-effect linear model, and a school-effect linear model. Four indices are considered: The Pseudo-Bayes Factor (PsBF) and three forms of Spiegelhalter's DIC that are named conditional DIC (DIC_c), group-marginalized DIC (DIC_g) and top-level marginalized DIC (DIC_t). To investigate the relative performance of these methods, three simulation studies are carried out. We evaluate the performance of the model selection indices for various conditions of IRT models, the number of individual-specific covariates, and the number of school-specific covariates. The results show that the models preferred by PsBF, DIC_g and DIC_t are almost coincident, and the conclusion from DIC_c is incompatible with those from the others. Considering computational viability and interpretability, DIC_g is the optimal index for MLIRT models. [29685]

114 An investigation of the differential math performance by country and beliefs about math and teaching using TEDS-M 2008
Mingying Zheng, *University of Nebraska - Lincoln, USA*

Although a large amount of research has been conducted on differential item functioning in testing, studies have focused on detecting differential item functioning rather than on explaining how or why it may occur. This study uses hierarchical generalized linear modeling to detect differential math knowledge of lower-secondary future math teachers due to cultural differences across three countries (U.S.A., Germany, and Singapore) and their corresponding beliefs about math and teaching, two variables that have been isolatedly examined in the previous literature in terms of mean performance as opposed to item performance. The relationship among item difficulty, country, and beliefs about math and teaching are also investigated using international data from TEDS-M 2008. [29437]

115 The difference in item scoring method under different culture background
Shan Zhiyan, *National Institute of Education Sciences, China*
Wang Weidong, *National Institute of Traditional Chinese Medicine, China*

In psychological testing, there is a general assumption that people with a higher trait level will get higher test scores when responding to test items. The cumulative response mechanism is the basis for traditional item scoring models (classical test theory, factor analysis model, and item response model). But in some situations, non-cumulative response models may be appropriate, especially when the subjects are from non-westernized cultures. In Confucian doctrine, moderation is the best of rules in all things, so the middle option of all alternatives may be the ideal one to choose. When Chinese subjects are responding to a 5-point Likert-type scale, the middle option (i.e., 3) is often considered as the best one rather than the first and the last ones (i.e., 1 and 5). So the middle option (other than the first or the last) should be assigned the highest score. The Guang'anmen Memory-tracing Personality Development Inventory (GMPI) is a localized scale which was constructed to assess Chinese personality in clinical settings base on Chinese medical theory and practice. Inspired by the scoring rules in the Generalized Graded Unfolding Model (GGUM), items in the GMPI were center-scored: the middle option was assigned the highest score, scores for other options decreased according to their distances from the middle point. The total score of GMPI was acquired after summing up all item scores according to classical test theory. The reliability and validity of GMPI had been verified. [29954]

116 Family context predictors of mathematics achievement among U.S. middle school students: A Bayesian hierarchical regression analysis
Hao Zhou, *University of Kentucky, USA*
Hongwei Yang, *University of Kentucky, USA*
Ruixue Liu, *University of Kentucky, USA*
Zijia Li, *University of Kentucky, USA*

This paper aims to examine the relationship between family context factors and middle school students' mathematics achievement in the US. To that end, a Bayesian hierarchical regression analysis is conducted where prior information is incorporated into the modeling process and independent variables are entered sequentially in the order of theoretical importance. By evaluating each model using Bayesian fit indices (say, BIC), a best and most

parsimonious model is selected where Bayesian statistical inference is conducted for the purpose of interpretation and discussion of results. The primary dataset used in this research is derived from the Program for International Student Assessment (PISA) conducted in 2012. A secondary PISA dataset from 2009 was analyzed under a traditional maximum likelihood regression analysis providing parameter estimates needed to specify informative priors for Bayesian modeling. The dependent variable in the study is literacy scores on mathematics. The independent variables include persons that students live with, mother's level of schooling, father's level of schooling, mother's job type, father's job type, numbers of books, numbers of electronic appliances, and numbers of educational resources at home. Finally, the entire analysis is performed using the Bayesian estimator provided in PROC GENMOD of SAS 9.4. [29769]

117 Studies of a method of controlling classification errors based on the analysis of item discrimination indices

Qun Zhou, *Shanghai Municipal Educational Examinations Authority, China*

According to the reform scheme of the Shanghai university entrance examination and enrollment system, the results of the Shanghai Senior High School Students Achievement Test will be used in selection of new students by the universities. It requires an achievement test with high reliability and low conditional standard errors at the cut scores. In China, however, the item development of large-scale educational assessment is conducted without trial tests and item selection, and the constructors assemble the test without item parameters for reference. Therefore, the test development process cannot ensure that the assessment is of high reliability and low conditional standard errors. By reexamining data from the General Social Science Test of Shanghai High School Achievement Examinations for the year 2013, this research tries to find a method of controlling classification errors based on the analysis of item discrimination indices. The author first employs discrimination indices to categorize valid items and invalid items, and then analyses how to adjust the test structures. It is found that the test should reduce the number of items for memory and comprehension abilities and add more items for testing synthesis and evaluation abilities, remove blank-filling items, construct more constructed-response items, and abolish the too difficult or easy items. The 2014 test results indicates that the modulations of the test structures on the basis of item discrimination indices could, to

some extent, increase the number of valid items and make the test structures in cognitive abilities, item types and difficulty coefficients more reasonable and thereby control the classification errors more effectively. [29782]

118 Assessing primary school students' learning on fraction by a concept map task based on cognitive diagnostic theory

Wanjing J. Zhou, *Beijing Normal University, China*
Jiaping P. Xie, *Beijing Normal University, China*
Tao Xin, *Beijing Normal University, China*

Based on Ausubel's meaningful learning, Novak and Ausubel developed a concept map, which is a kind of teaching and learning strategy, and a graphical tool organizing and displaying specific domain knowledge. The concept map can show the students' cognitive structures visually. From the 1960s to now, research on the concept map has become relatively mature, and results are quite abundant in the fields of psychology, education, computer science and others. In applications of the concept maps, assessing the students' learning is always a central issue. As a technology different from traditional testing methods, there are many studies that have already verified the suitability and unique role of the concept map in assessing students' learning. Concept maps have three main assessment tasks: A creative concept map task (C Technology), a selected concept map task (S Technology), and a filled concept map task (F technology). By reflecting on existing studies, we want to try a new assessment task, an answered concept map task. In the domain knowledge of fraction, combining concept map technology and modern psychological and educational measurement theory, experts defined and students drew concept maps to verify the key concepts, standard map, attributes and attribute hierarchy; designing the test according attribute hierarchy in cognitive diagnosis. So we can reflect the standard concept map by tests. We then apply rule space methodology to analyze a sample of 514 sixth-grade students' observed item responses to identify their cognitive structure. [29678]

119 The effects of autonomy, external motivation and self-efficacy on academic performance for high school students: A three wave longitudinal study

Yuan Zhou, *Beijing Normal University, China*
Hongyun Liu, *Beijing Normal University, China*

The study describes the mechanism of how learning autonomy, external motivation, and self-efficacy influence the academic performance of senior high school students,

based on longitudinal data on 24,559 students from 46 middle schools. In this study, the Latent Growth Model (LGM) with time constant covariates is used. Some important conclusions were obtained: first of all, the performance of new senior high school students is significantly influenced by their learning autonomy, external study motivation, and self-efficacy. Furthermore, it is suggested that high autonomy and external study motivation facilitate the developmental speed of academic achievement in new senior high school students. However, high self-efficacy was found to weaken the developmental speed of academic achievement in new senior high school students. Therefore, the cultivation of students' learning autonomy as well as providing external motivation for students should be emphasized. But there is an interesting phenomenon that excessive praise may not be so suitable for all students. [29860]

Author Index

Arai, Sayaka, 8

Barboza, Julissa, 8
Bian, Yufang, 39
Bo, Lü, 8
Bolt, Daniel, 21

Cai, Chao, 11
Cano-Vindel, Antonio, 13
Carlson, Ralph, 8
Carstensen, Claus H., 12, 30
Carter, Nathan T., 27
Chae, Hyun Jung, 8
Chang, Joyce, 35
Chen, Fu, 9
Chen, Pin-Wun, 9
Chen, Shu-Ying, 9, 10
Cho, Young Il, 9
Choi, Jinnie, 10
Chung, Seungeun, 9
Chung, Yoonkyung, 8
Clifton, James P., 10
Cohen, Allan S., 19
Cui, Zhongmin, 11

Dai, Hong, 35
Dai, Wugen, 13
Daus, Stephan, 11
de Rooij, Mark J., 20
Depaoli, Sarah, 10, 21, 41
Ding, Shu-Liang, 36
Ding, Shuliang, 27
Du, Jiaxuan, 17
Duncan, Ravit Golan, 10

Fairchild, Amanda J., 11
Fan, Xitao, 33
Fan, Yiping, 35
Fen, Luo, 32
Feng, Fan, 17, 37
Filus, Ania, 12
Fischer, Luise, 12, 30
Flores, Carlo Rafael, 8

Gao, Chunlei, 23
Gao, Fengqiang, 23, 38
Gao, Lingyun, 13
García Batista, Zoilo Emilio, 13
Gawlick, Lisa, 13
Guan, Kexin, 16
Guan, Li, 27
Guerra Peña, Kiero, 13

Guo, Fei, 14
Guo, Lei, 14
Guo, Miao, 38
Guo, Qingke, 14
Guo, Yage, 15
Guo, Yanfang, 9

Han, Jilin, 15
Han, Panpan, 37
Han, Woori, 15
Hao, Yi, 41
Hayashi, Kentaro, 24
He, Ye, 16
He, Yong, 11
Hong, Su Hyang, 9
Hu, Jie, 38
Huang, Lijie, 36
Huang, Xiaoting, 16
Huang, Yuheng, 43
Hutson, Bryant, 16

Jian, Gao, 16
Jiang, Yanxia, 17
Jianhua, Xiong, 17, 32
Jie, Qi, 31
Jo, Yujung, 9
Joung, Youn-Joon, 18
Ju, Unhee, 20

Kang, Chun-Hua, 34
Kang, Hyejin, 19
Kato, Kentaro, 18
Kawaguchi, Atsushi, 39
Kelecioğlu, Hülya, 20
Kim, HeeKyoung, 18
Kim, Insuk, 18
Kim, Jihyun, 9
Kim, Keonseob, 18
Kim, Kyoungshin, 18
Kim, Kyung Yong, 19
Kim, Meereem, 19
Kim, Nana, 19
Kim, Sewon, 20
Kim, Su-Young, 31
Kong, Feng, 36
Kuijpers, Renske E., 20
Kumlu, Gökhan, 20
Kuo, Bor-Chen, 22

Lai, Keke, 10, 21
Lee, Daniel, 40
Lee, Guemin, 9, 19, 40

Ward, Kimberly, 36
Watkins, Karen E., 18
Weidong, Wang, 44
Wen, Hong-Bo, 24, 25
Wenyi, Wang, 32
Wu, Yi-Jhen, 37

Xiang, Shibei, 37
Xie, Jiaping, 45
Xin, Tao, 9, 26, 29, 37, 41, 45
Xin, Xin, 38
Xing, Kuan, 38
Xiong, Jianhua, 27
Xu, Qinfang, 39

Yamamoto, Michio, 39
Yang, Chenchen, 39
Yang, Heewon, 40
Yang, Hongwei, 28, 43, 44
Yang, Ji Seung, 40
Yang, Tao, 23, 41
Yang, Xi, 40
Yang, Yuzhu (June), 21, 41
Yi, Qin, 23, 26, 41
Yi, Yumi, 15
Yin, Pei, 25
Yoo, Nayeon, 21
Yu, Chun, 10
Yu, Tian, 42
Yu, Xiaofeng, 23
Yuan, Jianlin, 42
Yuan, Ke-Hai, 24
Yue, Ma, 42
Yulong, Bian, 42
Yurttaş, Gulfem Dilek, 20

Zeng, Ping-Fei, 34
Zhang, Changhui, 13
Zhang, Danhui, 23
Zhang, Duan, 43
Zhang, Jinfu, 25
Zhang, Jinhua, 17, 37
Zhang, Jiwei, 35
Zhang, Qun, 43
Zhang, Xue, 44
Zhang, Yongmei, 41
Zhang, Yuxiao, 40
Zhao, Lin, 14
Zheng, Mingying, 44
Zher, Sheng Yee, 34
Zhiyan, Shan, 44
Zhou, Hao, 44

Zhou, Junli, 35
Zhou, Ning, 25
Zhou, Qun, 45
Zhou, Wanjing, 45
Zhou, Yuan, 45
Zhu, Haiyan, 25
Zhu, Xiaowen, 31
Zhu, Xinli, 38, 42