The 7th Conference of the International Test Commission

ITC

Challenges and Opportunities in
Testing and Assessment
in a Globalized Economy

July 19-21, 2010

July 18, 2010 (Pre-Conference Workshops)

The Chinese University of Hong Kong,
Shatin, Hong Kong
Website: http://www.itc2010hk.com
Email: itc2010@psy.cuhk.edu.hk

Table of Content

Welcome Message from Marise Ph. Born, President, International Test Commission

It is my great pleasure to welcome you on behalf of the International Test Commission and to invite you to attend the next International Test Commission conference to be hosted in Hong Kong in 2010.

The ITC forms an Association of national psychological associations, test commissions, publishers and other organizations, and is committed to promoting effective testing and assessment policies and to the proper development, evaluation and uses of educational and psychological instruments.

In the line of ITC conferences, the 2010 conference in Hong Kong is an exciting and definitely historic event, as the conference will be the very first to be held in a non-Western nation. The conference will be the 7th in a very successful series of ITC conferences which have become renowned for their strong focus and expertise in psychological and educational testing. The ITC conferences are known for their animated and personal atmosphere in which experts discuss and novices learn about cutting edge research and applications in this field of science. This enthusiastic mood perfectly fits the vitality and energy in East Asia, a region of the world which moves ahead in the domain of testing and assessment of individuals in a relentless pace.

I very much am looking forward to see you in Hong Kong in 2010.

Marise Ph. Born
President ITC

Welcome from the 2010 Conference Chair

It is my pleasure and privilege to invite you to join us at the 7th ITC Conference in Hong Kong. Testing and assessment is currently a fast emerging field in Asian countries. The historical roots of modern testing may be traced to the imperial examination system used to evaluate ability and select officials on the basis of merit in Han Dynasty China over 2000 years ago. This is the first time in its history that the ITC holds its conference in Asia. We hope that the 2010 conference will showcase cutting-edge developments in testing and assessment, and promote exchange on testing standards and guidelines in this part of the world.

As Asia's world city, Hong Kong is a culturally diverse and sophisticated metropolis that blends eastern and western influences into a dynamic destination. It offers a host of memorable tourist and shopping attractions within its compact area. You will be amazed by the diverse contrasts and close proximity of stunning cityscapes and soaring mountains, heritage sites and extensive green countryside. As a paradise for food lovers, Hong Kong is renowned as the culinary capital of Asia where you can taste the best regional and international cuisine.

The 7th ITC Conference is a satellite conference of the 2010 International Congress of Applied Psychology to be held in Melbourne. Delegates can conveniently combine the two conferences with a stopover in Hong Kong after the Melbourne conference.

The 7th ITC Conference will be held in the spacious campus of the Chinese University of Hong Kong overlooking the scenic Tolo Harbour, which is one of the largest and greenest areas in Hong Kong. The campus is located in Shatin, New Territories, with easy access to the city center via the well established transportation network.

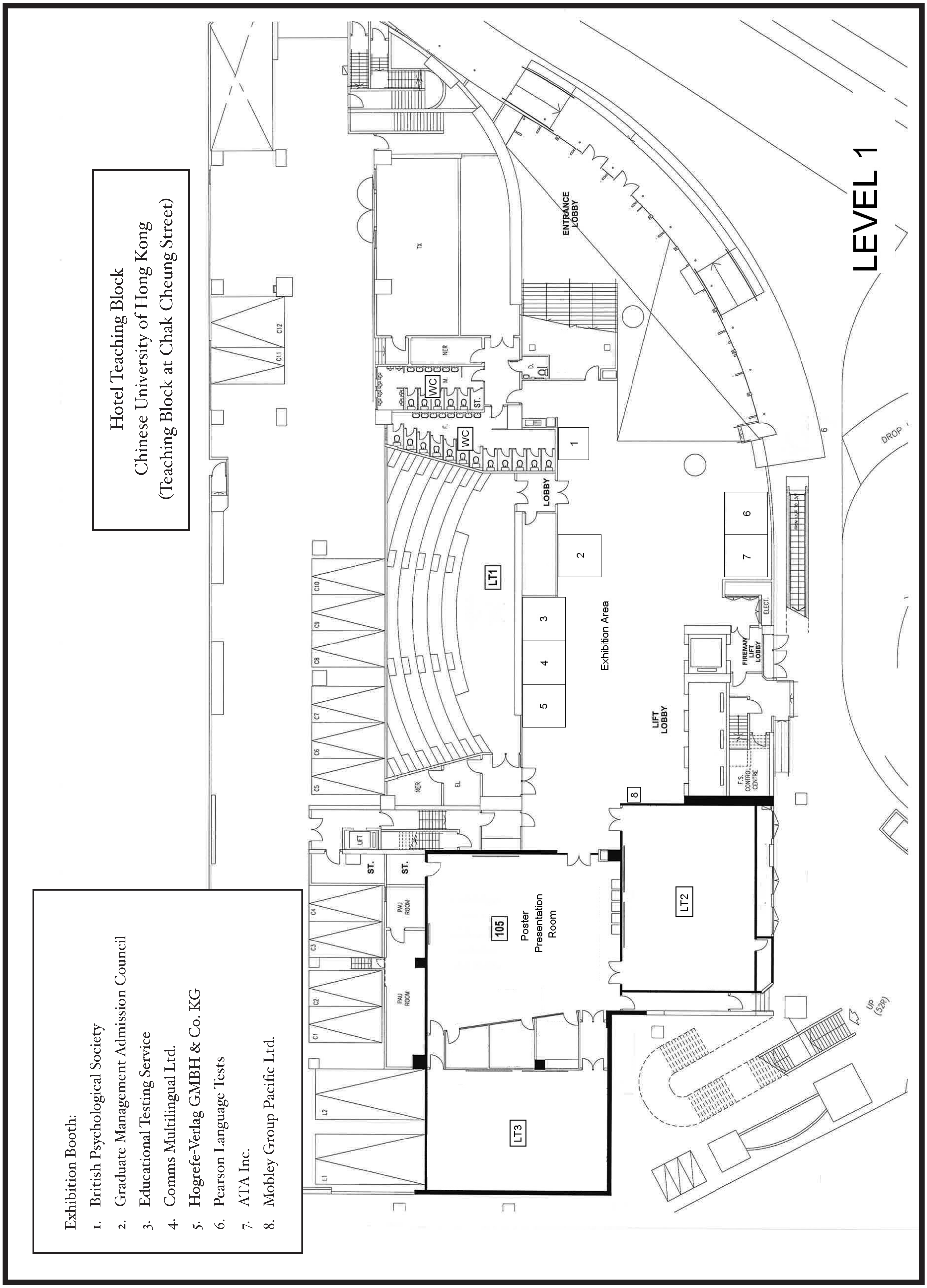
I look forward to meeting you at this exciting conference in the vibrant city of Hong Kong in 2010.

Prof. Fanny M. Cheung
Chair, Organizing Committee

Committees

Organizing Committee	Cheung, Fanny M. (Chair; The Chinese University of Hong Kong)
	Bartram, David (SHL Group)
	Byrne, Marise (Erasmus Unviersity Rotterdam)
	Byren, Barbara (University of Ottawa)
	Carey, Tim (The Chinese University of Hong Kong)
	Chan, Wai (The Chinese University of Hong Kong)
	Foster, David (KRYTERION)
	Fung, Helene (The Chinese University of Hong Kong)
	Grégoire, Jacques (Université Catholique de Louvain)
	Hau, Kit-tai (The Chinese University of Hong Kong)
	Leung, Freedom (The Chinese University of Hong Kong)
Scientific Committee	Born Marise (Chair; Erasmus University Rotterdam)
	Bartram, David (SHL Group)
	Cheung, Fanny M. (The Chinese University of Hong Kong)
	Grégoire, Jacques (Université Catholique de Louvain)
	Hambleton, Ronald (University of Massachusetts)
	Iliescu, Dragos (D&D Research)
International Advisory Panel	Bogg, Jan (University of Liverpool)
	Bullock, Merry (IUPsyS)
	Cheung, Mike (National University of Singapore)
	Foxcroft, Cheryl (Nelson Mandela Metropolitan University)
	Hattie, John (University of Auckland)
	He, Rongguei (Taiwan Normal University)
	Korah, Ringking M. (University of Indonesia)
	Laile, Jas S. J. (University of Malaya)
	Leong, Frederick T. L. (Michigan State University)
	Muñiz, José (University of Oviedo)
	Natarajan, V. (MeritTrac Services Pvt. Ltd)
	Nielsen, Sverre L. (The Norwegian Psychological Association)
	Oakland, Thomas (University of Florida)
	Oh, Kyung Ja (Yonsei University)
	Purwono, Urip (Padjadjaran University)
	Shigemasu, Kazuo (University of Tokyo)
	Yan, Gonggu (Beijing Normal University)
	Zhang, Jianxin (Chinese Academy of Sciences)

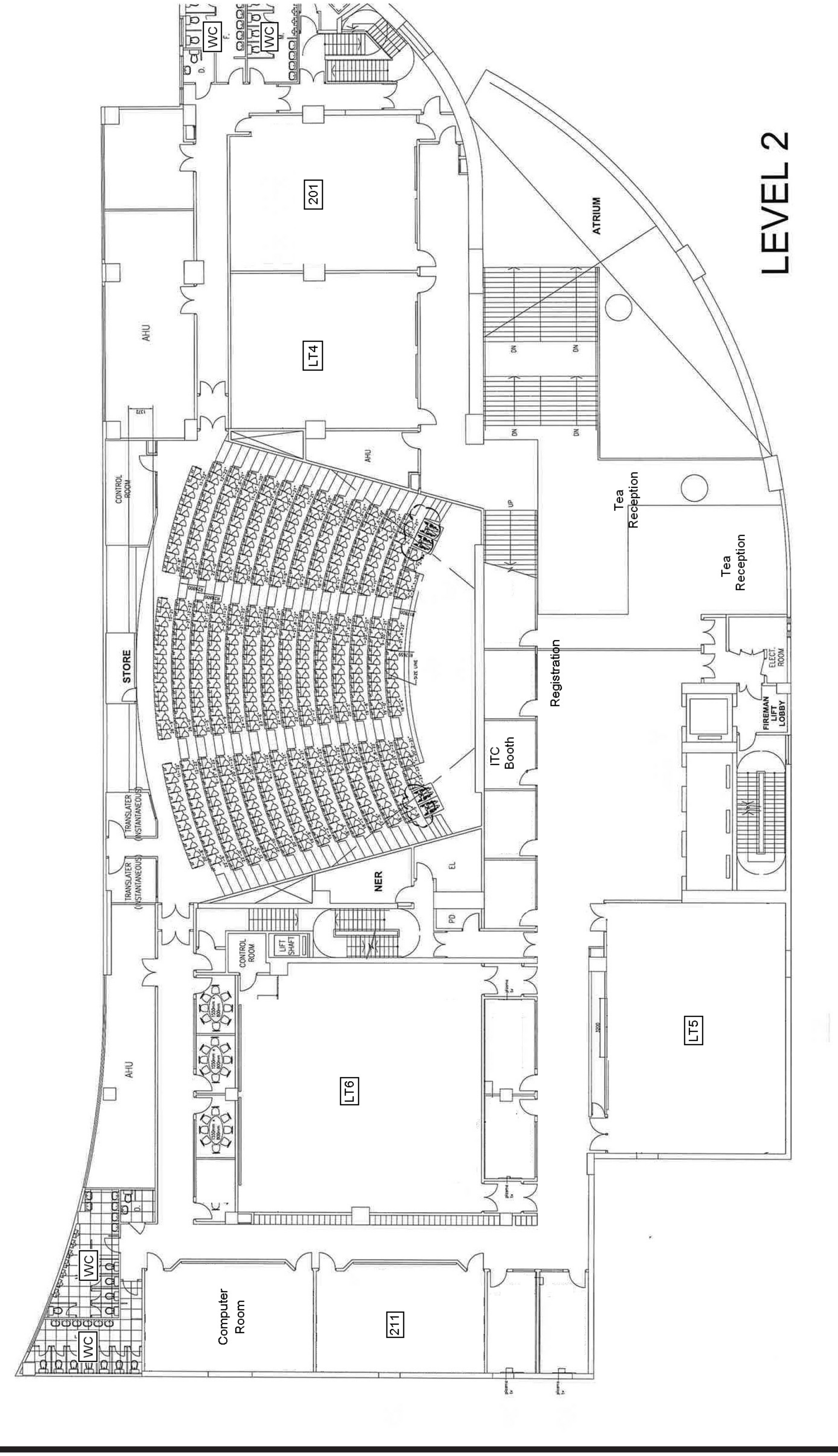
Floor Maps of Conference Venue

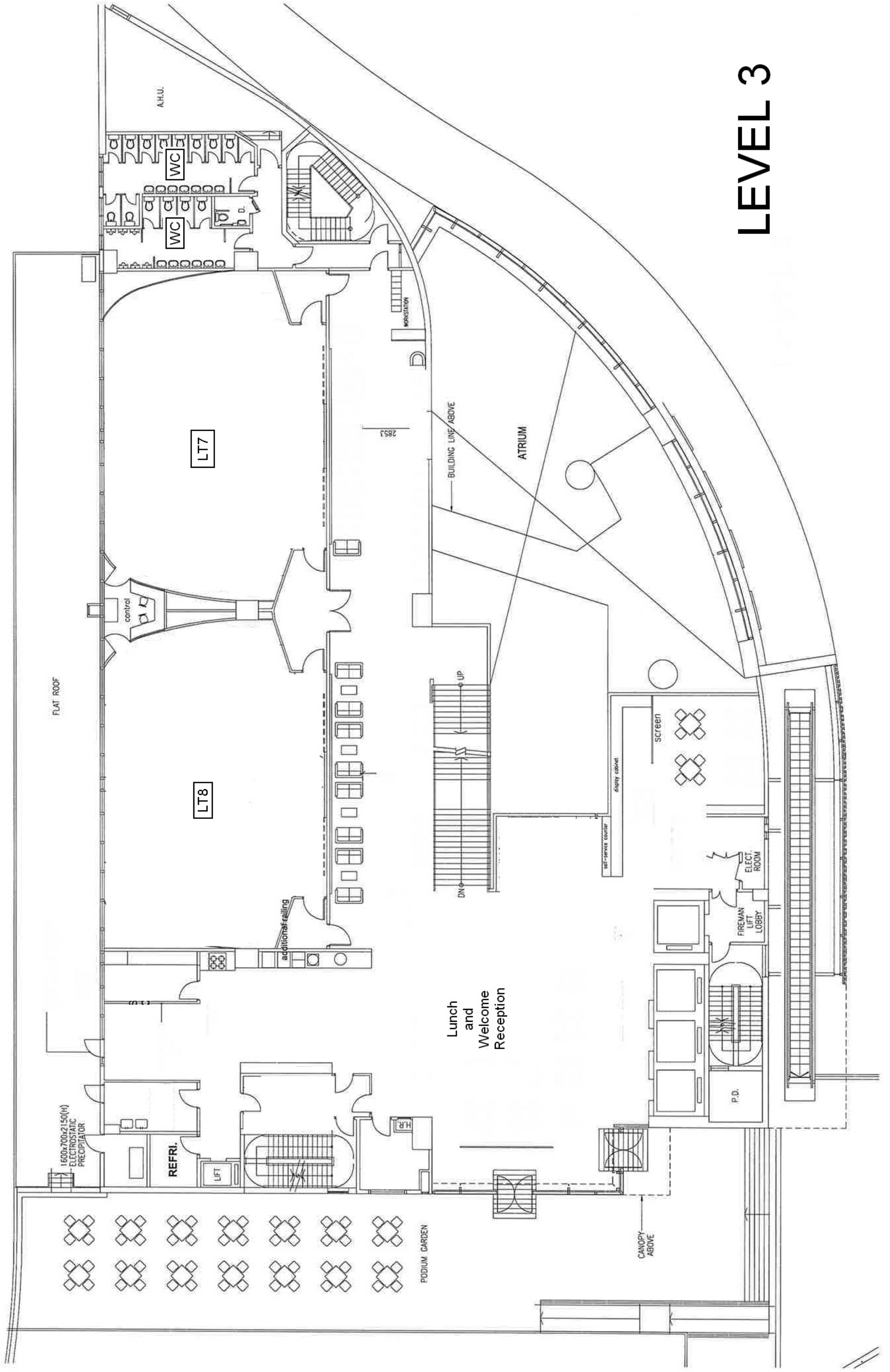


Hotel Teaching Block
Chinese University of Hong Kong
(Teaching Block at Chak Cheung Street)

Exhibition Booth:

1. British Psychological Society
2. Graduate Management Admission Council
3. Educational Testing Service
4. Comms Multilingual Ltd.
5. Hogrefe-Verlag GMBH & Co. KG
6. Pearson Language Tests
7. ATA Inc.
8. Mobley Group Pacific Ltd.


LEVEL 2
WC
WC
201
LT4
AHU
ATRIUM
DN
DN
DN
DN
AHU
UP
Tea Reception
Tea Reception
CONTROL ROOM
STORE
Registration
ITC Booth
ELECT. ROOM
FIREMAN LIFT LOBBY
TRANSLATER (INSTANTANEOUS)
TRANSLATER (INSTANTANEOUS)
NER
EL
PD
CONTROL ROOM
LIFT SHAFT
AHU
LT6
LT5
WC
WC
Computer Room
211


LEVEL 3
A.H.U.
WC
WC
D.
WORKSTATION
LT7
2853
BUILDING LINE ABOVE
ATRIUM
control
FLAT ROOF
LT8
UP
DN
screen
display cabinet
self-service counter
ELECT. ROOM
FIREMAN LIFT LOBBY
additional railing
Lunch and Welcome Reception
P.D.
H.R
LIFT
REFRI.
1600x200x2150(H) ELECTROSTATIC PRECIPITATOR
PODIUM GARDEN
CANOPY ABOVE

Scientific Programme

7th Conference of the International Test Commission

19-21 July, 2010

The Chinese University of Hong Kong
Shatin
Hong Kong

Main Theme	Challenges and Opportunities in Testing and Assessment in a Globalized Economy
Five sub-themes	Developments in psychometrics and test theory for international testing Indigenous, second language, and cross national test development Geotrends in testing: making use of technology advances in test administration and data management Issues of policy, ethics, professionalism and training in multinational testing Test security and privacy concerns when testing internati

Programme Schedule - Monday July 19, 2010

Time	LT1	LT2	LT3	LT4	LT5	LT6	Room 201	Room 211	Room 105
7:30-8:30	Registration (Level 2)								
8:30-9:15	Opening Ceremony (LT1)								
9:15-10:45	Plenary Session (LT1) State-of-the-art Speech by Robert Roe Speech by Dave Wilson from GMAC								
10:45-11:00	Tea Break								
11:00	***Keynote Address 1*** ***Validation support for selection procedures*** ***Schmitt, Neal*** Weiss, Larry (session chair)	***Invited symposium 1*** ***Allalouf, Avi*** Bennett, Randy Hambleton, Ronald Zhang, Houcan Grégoire, Jacques Gafni, Naomi		***C01*** ***Leong, Frederick*** Arce-Ferrer, Alvaro Li, Xuhang Schulruf, Boaz Wu, Joseph Zumbo, Bruno D.	***Special Session 1*** ***Debating the cost of psychological tests and the factors that determine the cost*** Foxcroft, Cheryl Bartram, Dave Foster, David Oakland, Tom	***Diamond Sponsor Session 1*** ***Anastasio, Ernie*** Guo, Fanmin Talento-Miller, Eileen Defibaugh, Courtney Taliaferro, Hillary Rudner, Lawrence		***S01*** ***Brown, Gavin T L*** Michaelides, Michalis Gao, Lingbiao Hui, Sammy K. F. Kennedy, Kerry J.	
11:30									
12:00									
12:30-14:00	Lunch								
14:00	***Keynote Address 2*** ***Ethical and other professional issues: What to do when working in the absence of local standards*** ***Oakland, Thomas*** Purwono, Urip (session chair)	***C03*** ***Talento-Miller, Eileen*** Clinton, Janet M. Zheng, Ying Johnston, Michael Orchard, Sue Zumbo, Bruno D. Michael, Joan J.		***C02*** ***Leung, Kwok*** Chen, Lijun Eatchel, Nikki Vrignaud, Pierre Wang, Y. Lawrence Xu, Jian Ping	***S02*** ***Weiner, John*** Burke, Eugene Fetzer, Michael	***Special Session 2*** ***International Test Commission Guidelines for adapting educational and psychological tests (2nd edition)*** Bartram, David Gregoire, Jacques Hambleton, Ronald Van de Vijver, Fons	***S04*** ***Zhou, Dan*** Li, Bo Li, Chongliang Li, Qi Xu, Jiehong		***P01*** Chao, Yu-Ning Gomiero, Tiziano Kreuzpointner, Ludwig Kuo, I-Ting Li, Ying Lu, Cheng-Chin Mitina, Olga Morley-Kirk, James Ni, Chen-Hao Preuss, Achim Roberts, Patricia Hill, Jill Gnambs, Timo Tono, Suwartono Hung, Pi-Hsia Sun, Jianan Wang, Bo
14:30									
15:00									
15:30-15:45	Tea Break								
15:45		***C04*** ***Schwarzer, Ralf*** Cheung, Shu Fai Chiou, Hawjeng Mitina, Olga Yuan, Ke-Hai Zhang, Zhiyong	***C05*** ***Wang, Wen Chung*** Albers, Frank Arce-Ferrer, Alvaro Chernyshenko, Oleksandr Huelmann, Gerrit Leung, Chi Keung Eddie Zierke, Oliver	***S06*** ***Leung, Freedom Yiu Kin*** You, Jianing Lai, Ching Man Fu, Kei	***Invited Symposium 2*** ***Nielsen, Sverre Leonard*** Lindsay, Geoff Oakland, Tom Bartram, Dave Hagås, Per Olav	***Invited Symposium 3*** ***Weekers, Anke*** Schmitt, Neal De Fruyt, Filip Serlie, Alec Van De Vijver, Fons	***C06*** ***Elosua, Paula*** Choi, Hye-Jeong Defibaugh, Courtney Solano-Flores, Guillermo Talento-Miller, Eileen Zheng, Ying	***S05*** ***Harris, William*** Schuchart, Nadine Tong, Alex Burke, Eugene	***P02*** Chen, Su-Yu Ding, Shu-Liang Fang, Ping Ghadimi Moghaddam, Malek Mohammad Kuo, Bor-Chen Kwok, Oi-Man Leeson, Heidi Leontiev, Dmitry Zhang, Minqiang Liu, Yan Lozano, Luis M. Lu, Szucheng Wang, Aijun Wang, Wen-Yi Xie, Qin Yang, Zhiming Yoon, Myeongsun Zheng, Rui Xin, Tao You, Xiaofeng
16:15									
16:45									
17:15-17:30	Break								
17:30-19:30	Welcome Reception (Level 3)								

Bolded names: Session chairs

Bolded names: Session chairs

Programme Schedule - Tuesday, July 20, 2010

Time	LT1	LT2	LT3	LT4	LT5	LT6	Room 201	Room 211	Room 105
8:00-8:30	Registration (Level 2)								
8:30			*S03* ***Cheung, Kwok Wah*** Cheung, Kwong Yuen, ThomasLam, Ling Chi, TennyTang, Mei Shin Chan, Ka Ki, Catherine	*C08* ***Roth, Hans*** Foxcroft, Cheryl Van De Vijver, Fons Yu, Guoxing Zhang, Sheng Hill, Jill	*S07* ***Fan, Weiqiao*** ***Zhou, Mingjie*** Wan, Sarah Lai Yin Cao, Hui	*S08* ***Chen, Po-Hsi*** Wu, Chia-Ju Chao, Hsiu-Yi Hsu, Chia-Ling Chen, Jyun-Hong	*C09* ***Nielsen, Sverre*** Bartram, Dave Phelps, Richard Preuss, Achim Serlie, Alec W. Sharf, James	*C07* ***Weekers, Anke*** Choi, Youn-Jeng Griffin, Patrick Kingston, Neal	*P03* Coscarelli, Alessandra Barbot, Baptiste Chang, Kenneth Charalampous, Kyriakos De Bastiani, Elisa Ding, Ching-Huei Elosua, Paula Fida, Kashif M. Iversen, Ole Lam, Ben C. P. Laurentiu P., Maricuoiu Lee, Leanda Lozano, Luis M. Megherbi, Hakima Okonkwo, Judith So, Timothy Vrignaud, Pierre Wan, Shih-Ting Wu, Shiu-Lien Brown, Gavin Thomas Lumsden
9:00	*Keynote Address 3* ***International high-stakes online testing: Best practices for test security and data privacy*** ***Foster, David*** Byrne, Barbara (session chair)	*Special Session 3* ***The international Journal of Testing: Ten Years and Going Strong*** Hattie, John							
9:30									
10:00-10:30		Tea Break							
10:30		*S09* ***Burke, Eugene*** ***Tong, Alex*** Liu, Ying	*S18* ***Bernstine, Daniel O.*** Van Tol, Joan Vaseleck, James Pashley, Peter Rudner, Lawrence	*S12* ***Gilmore, Alison*** Smith, Jeffrey K. Darr, Charles Smith, Lisa Hattie, John	*S10* ***von Davier, Alina A.*** Gorham, Jerry Eggen, Theo Yoo, Hanwook Fetzer, Michael S. Gafni, Naomi		*C10* ***Natarajan, Venkatesa*** Barrada, Juan Ramon Han, Kyung (Chris) T. Lin, Jyun-Ji Liu, Hongyun Walker, Cindy M.	*S11* ***Ding, Yi*** Kuo, Yi-Lung	*P04* Ertubey, Candan García-Rueda, Rebeca Gillet, Isabelle Kuo, Bor-Chen Hsu, Chun-Yu Lin, Yi-Hung Lozano, Luis M. Li, Tao Ding, Shu-Liang Molchanov, Alexander Okonkwo, Judith Primi, Caterina Rogers, H. Jane Shih, Pei-Chun Su, I-Hsiang Wang, Aijun Zhang, Minqiang Zhao, Yue
11:00	*Keynote Address 4* ***From indigenous to cross-cultural personality assessment: The usefulness of the combined emic-etic approach*** ***Cheung, Fanny M.*** Leong, Frederick (session chair)					*Special Session 4* ***Examining Formative Assessment*** Bennett, Randy Brown, Gavin Koh, Kim			
11:30									
12:00-13:30	Lunch								
13:30	*S19* ***Fremer, John*** Burke, Eugene Geranpayeh, Ardeshir Tong, Alex Sun, James Jian-Min	*S13* ***Wang, Wen Chung*** Lee, Kung-Hsien Huang, Sheng-Yun Chen, Po-Hsi	*S14* ***Chen, Shu-Ying*** Chen, Jyun-Hong Lin, Yi-Hung Liu, Tzu-Chen Lee, Hsiang-Ling	*S15* ***Gan, Yiqun*** Ng, Alexander Yao, Jingdan Fan, Weiqiao Leung, Kwok	*Invited Symposium 4* ***Purwono, Urip*** Halim, Magdalena Mansyur, Mansyur Salim, Djohan Mardhapi, Djemari Suhapti, Retno	*Diamond Sponsor Session 2* ***Schuchart, Nadine*** Gillet, Isabelle Li, Tao Roth, Hans J. Schuchart, Nadine	*C12* ***Ercikan, Kadriye*** Egeland, Jens Hill, Jill Malykh, Sergey Mordeno, Imelu Yan, Gonggu	*C11* ***Johnston, Michael*** Gessaroli, Marc Lee, John Chi-Kin Mamauag, Maria Felicitas (Marife) M. O'Neill, Thomas Woo, Ada	*P05* Curkovic, Natalija Dela Rosa, Elmer Ganotice, Fraide Li, Jian Li, Zhongquan Turilova-Miščenko, Tatjana Zulu, Happy Amurao, Analiza Liezl Arce-Ferrer, Alvaro Ding, Shu-Liang Gorham, Jerry Han, Yuna Lee, Pei-Yu Lin, Chien-Yu Molchanov, Alexander Park, Yoon Soo Cheng, Chien-Ming Wang, Li Jun Ye, Shengquan Yu, Jiayuan Ong, Saw Lan
14:00									
14:30									
15:00-15:15	Tea Break								
15:15	*Invited Symposium 5* ***Hambleton, Ronald K.*** Huff, Kristen Mills, Craig Zumbo, Bruno D.	*S16* ***Mok, Magdalena Mo Ching*** ***Wang, Wen Chung*** Wang, Li-Jun Tam, Hak-Ping Cheng, Rebecca Wing-Yi Lee, Tony	*C13* ***Han, Chris*** Barbot, Baptiste Care, Esther Chiou, Hawjeng Lee, Young-Sun Magno, Carlo Park, Yoon Soo	*Invited Symposium 7* ***Shigemasu, Kazuo*** Muraki, Eiji Mayekawa, Shinich Nogami, Yasuko Otsu, Tatsuo	*Invited Symposium 6* ***Fontaine, Johnny*** Wong, Chi-Sum Roberts, Richard Yik, Michelle Grégoire, Jacques	*S17* ***Bartram, Dave*** Evers, Arne Born, Marise Sun, James Jian-Min	*C16* ***Gilmore, Alison*** Alexeev, Natalia Ong, Saw Lan Phelps, Richard Shih, Shu-Chuan Syaifuddin, M.	*C14* ***Leung, Freedom*** Bittner, Jenny V. Dasari, Venkata Venu Gopal Gori, Alessio Leung, Cynthia Tao, Vivienne Y. K. Shahid, Mamoona	*P06* Besharat, Mohammad Ali Chang, Te-Sheng García Meraz, Melissa Hanfstingl, Barbara Hatami, Mohammad Hou, Ya-Ling Huang, Duan Hung, Su-Pin King, Ronnel Chang, Kuei-Lin Li, Xueyan Michaelides, Michalis Pariñas, Neil Vrignaud, Pierre Xu, Jian Ping Yang, Min Nacira, Zellal Zhang, Min-qiang
15:45									
16:15									
16:45			*C15* ***Chang, Lei*** Ambreen, Saima Bertling, Jonas Pablo Jeng, Hi-Lian Magno, Carlo Marberger, Tove Kanestrom	*S20* ***Cowieson, Neil*** Fraccaro, Michael U, Kin Chong Fung, Helen To, Clara			*C17* ***Mok, Magdalena Mo Ching*** Yang, Zhiming Han, Kyung (Chris) T. Kubinger, Klaus D Fung, Tze-Ho Huang, Hung-Yu Chernyshenko, Oleksandr		
17:15						*AGM of ITC and Town Hall Meeting*			
17:45									
18:15-18:30									
18:30-21:00	Banquet (The Star Seafood Floating Restaurant)								

Bolded names: Session chairs

Time	LT1	LT2	LT3	LT4	LT5	LT6	Room 201	Room 211	Room 105
8:00-8:30	Registration (Level 2)								
8:30		***S21*** ***To, Clara*** Bartram, Dave Elliott, Ray Sue-Chan, Christina Leung, Kwok	***S22*** ***Cheung, Shu Fai*** ***Born, Marise*** Leong, Frederick Iliescu, Dragos Dang, Minh	***C19*** ***Oakland, Tom*** Cheng, Christopher H K Freund, Alexander Mcinerney, Dennis Sava, Florin A. Hill, Jill Van Luijk, Frank	***Invited Symposium 8*** ***Weiss, Lawrence G.*** Chen, Hsin-Yi Li, Yuqiu Zhang, Houcan Zou, Yizhuang Grégoire, Jacques	***Invited Symposium 9*** ***Elosua, Paula*** ***Hambleton, Ronald K.*** De Boeck, Paul A.L. Elosua, Paula Zumbo, Bruno D.	***S23*** ***Burke, Eugene*** Harris, William G. Rudner, Lawrence Sun, James Jian-Min Geisinger, Kurt. F. Foster, David	***C18*** ***Weekers, Anke*** Dodeen, Hamzeh Zheng, Ying Kuo, Bor-Chen Swaminathan, Hariharan Wehrmaker, Maike	***P07*** Besharat, Mohammad Ali Curkovic, Natalija De Bastiani, Elisa Fan, Xiao Ling Gomiero, Tiziano King, Ronnel Leung, Beeto Lin, You Zhen Penelo, Eva Rohe, Anna Sundseth, Øyvind Wu, Chiao Ying Xu, Jian Ping Yang, Xin Sophie Ghamarani, Amir Zhang, Wenjing Lv, Shaobo
9:00									
9:30									
10:00	***Keynote Address 5*** ***Recent developments in international testing*** ***Van De Vijver, Fons*** Leong, Frederick (session chair)	***Special Session 5*** ***Informing about ISO 10667 - An International Standard for Assessment Service Delivery in Work andOrganizational Settings*** Born, Marise Bartram, Dave Nielsen, Sverre Geisinger, Kurt Tong, Alex Harris, William G.	***S24*** ***Wechsler, Solange*** Oakland, Thomas Hutz, Claudio Byrne, Barbara	***S25*** ***Chan, Agnes Sui-Yin*** Cheung, Mei-Chun Sze, Sophia, Lai-Man Chang, Sonia	***Invited Symposium 10*** ***Zhang, Jianxin*** Zhang, Minqiang Gan, Yiqun Huang, Zheng Wang, Li Zhang, Houcan	***Invited Symposium 11*** ***Hambleton, Ronald K.*** Wandall, Jakob Hamp-Lyons, Liz Hattie, John Ercikan, Kadriye von Davier, Alina A.		***C20*** ***Guo, Fanmin*** Ackermann, Kirsten Megherbi, Hakima Mulhern, Gerry Lee, Tony Ruiz-Primo, Maria Araceli Thomas Ahluwalia, Nancy	
10:30									
11:00									
11:30-11:45	Tea Break								
11:45-13:00	Closing Ceremony and Plenary Session (LT1) State-of-the-art Speech by John Hattie								

Bolded names: Session chairs

Social Event

Welcome Reception

All participants are warmly invited to attend the welcome reception, to be held on Level 3 of the Conference venue, Monday, 19th July 2010 at 5:30-7:30pm. Wine and snacks will be served.

Banquet

Date:	Tuesday 20 July 2010
Time:	6:30pm – 9:30pm
Venue:	The Star Seafood Floating Restaurant No. 55, Tai Chung Kiu Road, Shatin, NT, Hong Kong
Cost:	HK$500

Note:
Banquet participants please bring along your banquet ticket and gather at the Entrance Lobby on Level 1 at 6:30pm on 20 July We have arranged for coaches to bring you to the restaurant.

MORNING WORKSHOPS

18 July 2010 (Sunday) *9:00am – 12:30pm*

Workshop 1 Room 201

Introduction to structural equation modeling

Chan, Wai (Chinese University of Hong Kong, Hong Kong SAR, China)

Abstract

Structural equation modeling (SEM) is one of the most widely used statistical techniques in social and behavioral sciences. The purpose of this workshop is to provide participants with a basic introduction to the concepts, practices, and applications in SEM. In particular, we will cover topics including symbols and notations used in SEM, path analysis, confirmatory factor analysis, and models with latent variables. The Bentler-Weeks model will be explained and described in terms of its connection with EQS, a statistical software program in SEM. In this workshop, we will minimize our emphasis on technical or statistical details of SEM. It would, however, be helpful if participants have had some training and experience with linear regression analysis.

Interested participants may consider registering for another advanced SEM session which follows this introductory workshop, where participants will have an opportunity to learn how to apply SEM using EQS.

Workshop 2 LT5

Evaluating test quality as users and writing manuals as authors: Two sides of a coin

Geisinger, Kurt F. (Buros Center for Testing, University of Nebraska-Lincoln, USA)

Abstract

This workshop is aimed at two audiences: test developers and test users. Test developers need to provide potential users with specific information so that these individuals can decide if the characteristics of the test meet their needs. Similarly, test users must look for answers to specific questions when deciding on the tests to use. These questions include intended uses of the measure, test development procedures including fairness procedures and analyses, reliability and validity evidence, the availability of norms and other scoring concerns, whether multiple forms have been developed and equated, the skills needed for test administration and score interpretation, whether the test is available in different languages (and how such new forms were developed), and whether it is appropriate for individuals with disabilities. What information needs to be made available and what may be left confidential, and where to find this information will also be discussed.

Workshop 3 LT4

Methods and designs for enhancing cross-cultural invariance

Leung, Kwok (City University of Hong Kong, Hong Kong SAR, China)

Abstract

Cross-cultural studies are regarded as quasi-experimental research, and threats that jeopardize the validity of cross-cultural differences and their explanations are reviewed. The consilience approach is advocated for strengthening cross-cultural invariance, which calls for diverse evidence based on a sound theoretical basis, multiple sources of data, different research methods, and explicit refutation of alternative interpretations. Three broad strategies for strengthening cross-cultural invariance are proposed under the consilience framework, including the systematic contrast of cultural groups, the inclusion of covariates to rule out alternative explanations, and the use of multiple research methods.

Workshop 4 LT6

Psychometric methods for investigating differential item functioning (DIF) and test bias: Concepts, methods and applications

Zumbo, Bruno D. (University of British Columbia, Canada)

Abstract

Methods for detecting differential item functioning (DIF) and scale (or construct) equivalence typically are used in developing new measures, adapting existing measures, or validating test score inferences. DIF methods allow the judgment of whether items (and ultimately the test they constitute) function in the same manner for various groups of examinees, essentially flagging problematic items or tasks. In broad terms, this is a matter of measurement invariance; that is, does the test perform in the same manner for each group of examinees? You will be introduced to a variety of DIF methods, some developed by the presenter, for investigating item-level and scale-level (i.e., test-level) measurement invariance. The objective is to impart psychometric knowledge that will help enhance the fairness and equity of the inferences made from tests. Topics include: (a) What is measurement invariance, DIF, and scale-level invariance? (b) Construct versus item or scale equivalence (c) Description of DIF methods (d) Description of scale-level invariance, (e) Examples, and (f) Recommendations.

AFTERNOON WORKSHOPS
18 July 2010 (Sunday) 1:30am – 5:00pm

Workshop 5 LT5

Establishing the ITC Guidelines on quality control in scoring, analysis and reporting of test scores

Allalouf, Avi (National Institute for Testing and Evaluation, Israel)

Abstract

In scoring, test analysis and the reporting of test scores, accuracy is essential. An inaccurate score resulting from wrong judgment, incorrect conversion of raw scores to standard scores, or accidental reporting of scores to the wrong client, are all examples of mistakes that should not occur.

Since 2008, the ITC has been developing a new set of Quality Control (QC) Guidelines for all types of measurement – psychological, educational and occupational. A draft of the Guidelines, constructed by Avi Allalouf with the help of Marise Born was distributed to eleven experts from various disciplines and different countries. They did tremendous work and their comments were of great value in revising the draft.

In the workshop the current version of the QC Guidelines will be presented and explained. Then, errors that might occur at each stage will be discussed, as well as examples and QC procedures for avoiding, detecting or correcting these mistakes. Models that deal with the causes of human error and ways to predict and reduce error will also be presented. Participants will be given hands-on practice in detecting various types of errors.

Workshop 7 Room 201

Testing basic structural equation models: Overview and hands-on application using the EQS approach

Byrne, Barbara (University of Ottawa, Canada)

Abstract

This workshop details the many stages of structural equation modeling (SEM) analyses and provides for hands-on application based on the EQS program (PC version). Following an overview of program notation and review of procedures involved in testing for the validity of hypothesized SEM models, participants are "walked through" both the specification of and results derived from the testing of two confirmatory factor analytic (CFA) and one full SEM model. Although data and software will be provided, workshop participants are required to bring their own laptops. A basic understanding of both factor analysis and SEM is a necessary prerequisite.

Workshop 8 LT6

Item Response Theory: Introduction to concepts, models, parameter estimation and fit, and several applications

Hambleton, Ronald K. (University of Massachusetts at Amherst, USA)

Abstract

Many testing agencies and researchers would like to use item response theory (IRT) models for developing, scoring, identifying bias, and equating of their aptitude, achievement, and personality tests. These IRT models, too, can be used to provide the measurement underpinnings for new test designs such as multi-stage testing and computer-adaptive testing. In this workshop, we will survey the following topics and provide several examples and practical experiences:

- Shortcomings of classical test theory that have inspired the development of IRT models, and basic classical test theoretic concepts such as reliability and item analysis,
- Specific IRT models for fitting binary and polytomously-scored data (e.g., 1-, 2-, and 3-parameter logistic models, graded response model),
- Basics of item and ability parameter estimation,
- Graphical and statistical approaches for assessing model fit (e.g., RESID PLOTS-2),
- Introduction to IRT software (e.g., BILOG-MG, PARSCALE),
- Development of tests using item and test and target information functions, and relative efficiency,
- Computer-based testing: Issues, designs, item exposure, and advantages and disadvantages,
- Identification of potentially biased test items due to culture, content, translation, and other factors,
- Follow-up readings and research.

21st July, 2010, 11:45–13:00, LT1

Global testing, global opportunities, global challenges, and a global future for assessment

Hattie, John (University of Auckland, New Zealand)

Abstract

This session will provide a retrospect on contributions from this conference, promote major issues confronting the world of testing and measurement, provide some challenges to be confronted, and suggest an agenda for this future.

19th July, 2010, 09:15–10:45, LT1

Testing for travelers: Past and future

Roe, Robert A. (Maastricht University, The Netherlands)

Abstract

Unlike testing in other fields of science, psychological testing is essentially comparative. The prevailing technology of testing is based on the paradigm of "individual differences", which assumes that people are similar except for attributes singled out for comparison. The comparative approach to testing has worked well in homogeneous and stable communities where people spoke the same language, had the same social background and shared the same culture, that is, in well bounded regions with low social and geographic mobility. But in a globalizing world, where people continually travel and communities are poorly bounded, heterogeneous and changing, it seems to be less effective. There are two main problems: (1) the comparisons produced by tests are ambiguous as the scores reflect other sources of variation (e.g. demographics, culture, language, and time); (2) competing instruments for testing travelers may give different results and it is unclear which test (e.g. which publisher, country, language and date of creation) can best be used. How can these problems be resolved? This keynote argues that psychometrics as we currently know it is unlikely to provide effective solutions. Therefore, it proposes a change in perspective that might lead to another way of testing. Starting from a historical look at how psychological testing has developed in a global environment characterized by diversity and inequality, it highlights the parties involved in testing. It claims that recognition of multiple interacting actors, with their diverging roles, views, and interests, may on the one hand reveal conflict but on the other provide a basis for developing a novel paradigm in which tests get a new purpose and format better suited for the global world of travelers.

20th July, 2010, 11:00-12:00, LT1

From indigenous to cross-cultural personality assessment: the usefulness of the combined emic-etic approach

Cheung, Fanny M. (Chinese University of Hong Kong, Hong Kong SAR, China)

Abstract

Learning from the experience of adapting imported measures and following international guidelines in test development, cross-cultural psychologists have developed indigenous instruments that capture important dimensions of personality for the local cultural contexts. In this address, I will illustrate the combined emic-etic approach using the program of research involved in developing the Chinese Personality Assessment Inventory (CPAI). Incorporating indigenous and universal dimensions provides an opportunity to explore the universal and culture-specific dimensions of personality. These dynamic exchanges in cross-cultural personality assessment confront the challenges of "intellectual imperialism" in adopting translated measures and re-examine the controversy of the universality versus cultural specificity of personality structure.

20th July, 2010, 09:00-10:00, LT1

International high-stakes online testing: best practices for test security and data privacy

Foster, David (Kryterion, Inc.)

Abstract

More than ever it is important for organizations to measure the skills, talents and knowledge of people worldwide. Many certification programs, such as those in the information technology, medical and financial industries, are global in scope. University and college admissions programs receive applications from around the world. Pre-employment screening exams are used by companies to recruit potential employees at a worldwide level. Online universities and colleges are offering need to evaluate their students' knowledge regardless of where they live. The mobility of the worldwide workforce and student population, and the use of the Internet for marketing, communication, education, and assessment are recent and important factors which support these trends. As these tests and assessments lead to important or high-stakes decisions that affect the lives of individuals it is important that they are psychometrically sound, administered securely, and protect the privacy of examinees. The latter two are the subject of this keynote address. It will address several important questions. What are the more critical security risks when testing globally? What can be done today to reduce these risks? What promising new innovations in security are on the horizon? What are the specific data privacy issues to consider when providing a global testing program and enforcing critical security rules.

19th July, 2010, 14:00-15:00, LT1

Ethical and other professional issues: what to do when working in the absence of local standards

Oakland, Thomas (University of Florida, USA)

Abstract

Psychologists and other professionals recognize the need to know and adhere to the ethics code in the country or countries in which they work. However, most countries do not have ethics codes that govern the work of psychologists. Thus, psychologists working in countries that do not have an ethics code face a dilemma: they need to behave ethically yet do not know the guidelines or standards that govern these behaviors. Some cross-national conditions about which psychologists should be aware when working cross-nationally, especially in countries that may lack an ethics code, are discussed. These include knowledge of the host country's prevailing moral values, its laws and administrative policies, and ethics codes as well as policies approved by international agencies and associations. Eight guidelines applicable to test use are provided for psychologists working in host countries that lack ethics codes.

19th July, 2010, 11:00-12:00, LT1

Validation support for selection procedures

Schmitt Neal (Michigan State University, USA)

Abstract

Various means of supporting the use of selection procedures will be briefly summarized. The preponderance of the evidence on validity comes from criterion-related validation studies conducted over the last century primarily in the U.S. and Europe and very usefully summarized in meta-analyses. These meta-analyses indicate that measures of many constructs have practically meaningful implications for organizations and individuals. We also discuss the limitations of the primary data base that is the basis of these meta-analyses and propose a collaborative longitudinal data collection effort that would involve multiple organizations in various countries to address these limitations.

21st July, 2010, 10:00–11:00, LT1

Recent developments in international testing

van de Vijver, Fons J. R. (Tilburg University, the Netherlands and North-West University, South Africa)

Abstract

Cross-cultural aspects of psychological tests are increasingly important; for example, tests are adapted for use in new settings, assessment is conducted in multicultural groups, and recruitment is more and more done from an international applicant pool. The presentation will describe recent developments in cross-cultural assessment that are relevant for such applications. The topics are:

- beyond emic and etic measurement: toward a balanced treatment of culture in assessment;
- recent developments in test translations and adaptations;
- cultural loadings in assessment;
- recruiting for a global economy;
- bias and invariance testing.

Plenary Session Speaker

19th July, 2010, 09:15–10:45, LT1

Equity Interest in Testing and Measurement from an Interested Observer

Wilson, David A. (Graduate Management Admission Council, USA)

Abstract

There are those today who would argue that, indeed, the world is not improved by the measurement of competence or by psychological testing. They resist the accountability that is concomitant with measurement and decry the establishment of standards of performance.

Others demean the science of psychometrics and psychological testing through the creation and aggressive positioning of ill-conceived or poorly designed assessments.

And yet, in the face of these assaults on the profession, there seems to be little initiative on the part of the leadership to take up the charge; to speak with a strong and public voice about the need for standards and about the dangers inherent in incompetent measurement. If we lack the courage and resources to take up this challenge, we run the risk of indeed being irrelevant.

Special Sessions

Monday, 19th July 2010

Special session 1 11:00-12:00 LT5

Debating the cost of psychological tests and the factors that determine the cost

Foxcroft, Cheryl (Nelson Mandela Metropolitan University, South Africa)
Bartram, Dave (SHL Group Ltd, United Kingdom)
Foster, David (Kryterion, Inc., USA)
Oakland, Tom (Department of Educational Psychology, University of Florida, USA)

In many countries, test users and practitioners have raised concerns about the price of psychological and educational tests. Especially in countries with emerging economies the high cost of tests impacts negatively on good assessment practices (e.g., pirating and test users not purchasing the original versions from publishers). A balance needs to be achieved between raising sufficient funds for the further research and development of tests through test sales while also ensuring that practitioners who need to use the tests can in fact access them. This panel discussion will explore the factors related to determining the price of tests, and debate whether we are achieving the balance between getting enough revenue from test sales to research and develop tests but still keeping the price affordable for test users.

Special session 2 14:00-15:00 LT6

International Test Commission Guidelines for adapting educational and psychological tests (2nd edition)

Bartram, David (SHL Group, England)
Gregoire, Jacques (University of Louvain, Belgium)
Hambleton, Ronald (University of Massachusetts, Amherst, USA)
van de Vijver, Fons (University of Tilburg, the Netherlands)

Interest has been growing for years in the topic of translating and adapting educational and psychological tests from one language and culture to others. Today, many of the popular intelligence and personality tests are translated and adapted into 50 or more languages; achievement tests such as those used in the large scale international assessments like TIMSS and PISA are translated and adapted into more than 30 languages and cultures. In the United States, as one approach for handling cultural diversity and accommodations, many states are making their achievement tests available to students in more than a single language. Tremendous progress too has been made in the methodology for translating and adapting tests. The journals are publishing many articles on topics such as conducting judgmental reviews, studying construct equivalence, and identifying item level bias. The purpose of this session is to introduce the second edition of the International Test Commission (ITC) Test Adaptation Guidelines. The participants will discuss the 17 new Guidelines, the process used to develop them over the past three years, comparisons between the first and second edition of the Guidelines, and their relevance for test adaptation practices in cross-cultural assessments. The four participants in this presentation are part of the six person ITC committee who had the responsibility for producing the second edition of the Guidelines.

Tuesday, 20th July 2010

Special session 3 09:00-10:00 LT2

The International Journal of Testing: Ten years and going strong

Hattie, John (University of Auckland, New Zealand)

The status and purposes of the Journal will be outlined, and the notion of "international" and "testing" outlined. There will be a presentation on behalf of the current editors (Steve Sireci and Rob Meijer) and previous editors will talk about their experiences and answer questions.

Special session 4 11:00-12:00 LT6

Examining formative assessment

Bennett, Randy (Educational Testing Service)
Brown, Gavin (The Hong Kong Institute of Education, Hong Kong)
Koh, Kim (Nanyang Technological University, Singapore)

In elementary and secondary education, formative assessment is in vogue. A key reason for the popularity of formative assessment is, undoubtedly, the claims that have been made for its effectiveness. This session reviews the evidentiary sources cited for these claims, and summarizes how the effectiveness of formative assessment might be responsibly represented. Two discussants will react, including in their comments the results of their own research on teachers' beliefs and their formative assessment practices, and on the potential emotional and social consequences students may experience through self- and peer-assessment activities.

Wednesday, 21st July 2010

Special session 5 10:00-11:00 LT2

Informing about ISO 10667 - An International Standard for Assessment Service Delivery in work and organizational settings

Born, Marise (Erasmus University Rotterdam, the Netherlands)
Bartram, Dave (SHL Group)
Nielsen, Sverre (The Norwegian Psychological Association, Norway)
Geisinger, Kurt (Buros Center on Testing & University of Nebraska, USA)
Tong, Alex (ATA Inc., China)
Harris, William G. (Association of Test Publishers, UAS)

In March 2007, in Berlin, the Deutsche Industrielle Norm (DIN) took the initiative to develop an International Standard for Assessment Service Delivery, known as ISO 10667. This Standard is process-oriented and focuses on procedures and methods to assess people in work and organizational settings. In the Standard, a pre-assessment phase, an assessment delivery phase and a post-assessment review phase are distinguished. International Standards are developed by a process of several formal stages. In July 2010, ISO 10667 has entered the Enquiry stage. In this stage, the draft International Standard (DIS) is commented upon by ISO member bodies. Within this session, we offer a briefing on the content of the Standard and the stage its development is in. A panel of participants in ISO 10667 will subsequently discuss the Standard. The audience will be given the opportunity to look at the draft International Standard (DIS).

Monday, 19th July 2010

Diamond sponsor session 1 11:00-12:30 LT6

More than scores

Chair

Anastasio, Ernest J. (Graduate Management Admission Council, USA)

Symposium Abstract

Test taker databases include vast amounts of valuable information in addition to test scores. Demographic data and information about where examinees send their scores, for example, can provide actionable, current insight into the global marketplace. This session will discuss ways the test sponsor can collect, organize, and disseminate information that can be gleaned from test databases using the Graduate Management Admission Test® (GMAT®) as an example. First introduced 55 years ago, more than 50 percent of all GMAT® examinees today are non-US citizens. Annual publications such as the GMAT® Candidate Profile and Geographic Trend Reports (World, Asian, and European) as well as individualized reports and validity studies will be highlighted. During this session we will discuss how each of these services were designed and implemented to more efficiently reach our increasingly global client needs.

Paper 1

"Reporting examinee population demographic changes"

Guo, Fanmin (Graduate Management Admission Council, USA)*

Changes have been occurring in the examinee population of many large-scale assessments for many years. Technological and psychometric advances have allowed the expansion of test use around the world and an increasingly diverse demographic of examinees. This presentation discusses two topics in light of such changes. The first is how to report the demographic changes using five-year rolling data for profiling. In addition to a printed copy of the profile report which is distributed to GMAT using schools every year, an interactive web-based version is provided to make more detail available to Graduate Management Admission Council (GMAC®) member schools. Details of both forms of this report will be presented. The second topic describes the redesign of the printed five-year profiles to switch focus from the needs of US schools to that of schools worldwide. Some demographic variables were redefined with an international perspective and others have been removed or reformatted in the profile to fit our new global client base. The process used to identify the changes needed to be more responsive to the global marketplace, including the design and implementation of a global survey, will be discussed. The presentation will be facilitated using examples from the printed and interactive profiles.

Paper 2

"Updating validity data collection and management"

Talento-Miller, Eileen(Graduate Management Admission Council, USA)*

Although test sponsors are keenly aware of the need for continuing evidence for the validity of test scores for their intended use in various populations and situations, the users themselves often do not feel as pressured to study their own use of scores, trusting in the quality of the test based on previous research or non-statistical evidence. In the case of admission tests, helping schools understand the validity of scores for their specific current programs benefits not only the users but the test sponsors as well, expanding the evidence available and keeping up to date with changes that may stem from changing demographics in the examinee population. Continuing validity evidence is necessary to ensure the efficacy of scores for admission to global business programs. This presentation will elucidate how the sponsors of the Graduate Management Admission Test (GMAT) exam have updated the Validity Study Service to streamline data collection and management and improve the quality of the reports to ensure value for the users. In addition, the presentation will discuss the methodology behind various meta-analyses that have been conducted to summarize the data from diverse programs around the world.

Paper 3

"Increasing options: Communicating shifts in interest in international programs"

Defibaugh, Courtney (Graduate Management Admission Council, USA)*

As with many testing programs, the population of Graduate Management Admission Test (GMAT) examinees has been evolving over the years. As such, demands from the users of the GMAT have increasingly shown an interest in test takers outside of the United States asking such questions as "Who are they?" "Where do they want to study?" To reach this demand a new set of reports, Geographic Trend Reports, was developed by Graduate Management Admission Council (GMAC). The Geographic Trends series currently consists of four reports; World, Asian, European, and North American. The World Geographic Trend Report reveals score sending patterns for citizens of 10 regions of the world. Drilling down farther, the Asian and European Trend Report provides data analyses on the top 10 citizenship groups for the Asian and European World Regions. The North American Trend Report shows trends on those tests taken within the United States and Canada. This session will cover how the Geographic Trend Report series was designed and implemented to meet the needs of GMAC clients.

Paper 4

"Handling information requests: How and why"

Taliaferro, Hillary (Graduate Management Admission Council, USA)*

With background questions and historical data linked to test scores, testing companies have a resource that, if properly utilized, can be enormously helpful to test users. While the Graduate Management Admission Council (GMAC) produces a wide range of publically available publications, some school users desire more detailed and specific data analyses to understand their own marketplace—who the examinees are and where they are interested in studying—and help them to develop marketing strategies. Through these special information requests, data can be aggregated to focus information to better meet the needs of our global clients. A simple information request may require only mean total scores for a small group of examinees, such as students near a particular city. Responses to more complex requests may consist of numerous comparative graphs and tables. This session will highlight the types of questions and the processes used to complete both simple and complex information requests received from our clients around the world. Examples will show end products used in presentations, research projects, and marketing. Going beyond basic scores and standard publications helps build relationships and underscores that an organization is truly client oriented.

Discussant

Rudner, Lawrence (Graduate Management Admission Council)*

Discussant will summarize symposium session. The discussant will describe possible applications at an international level for products described during the session.

Tuesday, 20th July 2010

Diamond sponsor session 2 13:30-15:00 LT6

Testing across cultures

Chair

Schuchart, Nadine (Hogrefe Verlag GmbH & Co. KG, Germany)

Symposium Abstract

The drive to improve methodology for achieving cross-cultural equivalence when adapting psychometric tests has never been more important. In a world of increasing cultural diversity both across and within nations, the task of ensuring sensible interpretation and fair comparisons has grown in its complexity. Add to this the emergence of new testing formats and an evolving number of constructs of interest; the achievement of a solution to culture free testing seems to be ever more like a minefield. Indeed the first paper in this symposium questions the feasibility of achieving a solution at all and warns against achieving equivalence at the expense of the construct being measured. The second paper compares the functioning of a work based personality questionnaire across a number of European cultures giving special attention to item functioning rather than restricting the study to scores at the overall scale level. The third paper considers the special issues in adapting the same work based questionnaire to China; a culture very different from those across Europe. Our final paper presents findings from the cross-cultural adaptation of a Leadership Judgment questionnaire comparing the generalizability of leadership style preferences, leadership judgment and item difficulty level across a range of European countries.

Paper 1

"How to test the same thing in different cultures?"

Gillet, Isabelle (Editions Hogrefe France S.A.S., France)*

In the I/O field, the construction of scales to be deployed in multiple cultures/languages is a key issue for tests developers. ITC guidelines propose steps, in order to guarantee fairness, equity and equivalence across national versions. Tests users demand that different language versions have the same scales, the same number of items and the same items thinking this proves the equivalence of testing across countries; as if "people are all the same, all around the world,." This paper addresses some specific questions regarding the dangers of "smoothing" away real difference to achieve surface equivalence such as:

- "Missing" the country specific aspects of how constructs manifest in behavioural terms
- Throwing out genuine variance when we get rid of a scale which is not strictly equivalent in each country?
- Dropping items which prove to be very good in one country but not in another one?

The fact is that tests assess human beings; that human beings express themselves through language, values and behaviors (all aspects greatly impacted by culture); so adaptation of tests must therefore take account of this important reality... culture makes a difference.

Paper 2

"Measurement and structural equivalence of European versions of a work-based personality questionnaire"

Li, Tao (Hogrefe Ltd., UK)*

Measurement equivalence is an indispensable requirement for valid cross-cultural comparisons. This paper explores the adaptation of "The Business-focused Inventory of Personality" (BIP) to a range of European cultures. The BIP has been selected as the focus for the paper because it was originally developed in Germany. It is somewhat unusual to have the source language for test adaptation being anything other than English (most commonly US English but more recently British English too). The comparative analysis across the different national versions has, in this study, been carried out at an item level as well as the usual mean score comparison level. This is considered to be important since differential item functioning may be masked when only overall scale scores are compared; thus important information for cross cultural comparison can be obscured. The paper demonstrates the application of multiple indicators, multiple causes (MIMIC) structural equation model, a relatively new technique, in detecting differential item function.

Paper 3

"Adapting a European work based personality test to China"

Benoit, Andreas (Benoit Consulting, Hong Kong SAR, China)
Roth, Hans J. (Swiss Consul General in Hong Kong, Hong Kong SAR, China)*

In an ever more interconnected world, with economic aspects of life increasingly linked to socio-economic development, organizations will need to intensify the development of talent in the workforce in order to keep pace with the increasingly intense competition. In order to measure and further develop leadership and other stylistic competences, there is a need for top level diagnostic tools specifically adapted for the Chinese market to identify current individual strengths and weaknesses. The BIP is a work focussed personality test developed in Europe. BIP – "made in China" – will be tailored to the specific culture and norm-groups of China and will serve the particularly strong and growing need for talent development in this part of the world. This paper will focus on:

- Cultural differences and peculiarities which need to be considered when adapting the tool
- Procedures for translating into Mandarin and for data collection
- Presentation of the first results and feedback from Chinese users and professionals

Paper 4

"The development of a situational judgment test from a cross-cultural perspective"

Li, Tao (Hogrefe Ltd., UK)*

Situational Judgement Tests (SJTs) are increasingly popular but there are less frequent reports on the specific considerations arising from such tests regarding international adaptation. This paper will focus on what has been learned from adapting a British SJT, 'the Leadership Judgement Indicator' (LJI) to five European countries. The LJI is of particular interest in terms of the adaptation process, because in addition to measuring how effectively leaders judge the style most appropriate for the situation, there is also a measure of preferred style. This allows a comparison of the cross-cultural generalisability of each type of construct when measured in an SJT format. The LJI items are scored for judgement according to goodness of fit to a theoretical model such that the difficulty level of items is a further comparison of interest cross-culturally. Finally, given the model on which the LJI is based arises from Western leadership theory a question is raised as to the impact of this on generalisability of the test to Eastern cultures.

Discussant

Schuchart, Nadine (Hogrefe Verlag GmbH & Co. KG, Germany)*

Invited Symposia

Monday, 19th July 2010

Invited symposium 1 11:00-12:30 LT2

Exhibition on testing & measurement

Organizer and Moderator

Allalouf, Avi (National Institute for Testing and Evaluation, Israel)

Participants

Allalouf, Avi (National Institute for Testing and Evaluation, Israel)
Bennett, Randy (Educational Testing Service, USA)
Hambleton, Ronald (University of Massachusetts, USA)
Zhang, Houcan (Beijing Normal University, China)
Grégoire, Jacques (Université Catholique de Louvain, Belgium)
Gafni, Naomi (National Institute for Testing and Evaluation, Israel)

Abstract

A scientific exhibition on testing & measurement is currently being developed. NITE (National Institute for Testing & Evaluation) and the Bloomfield Science Museum, Jerusalem have already begun working on the scientific exhibits. The current partners are ETS (Educational Testing Service) and the Franklin Institute, Philadelphia. Among the topics presented by the exhibition will be: the history of testing and its role in society, reliability, validity, intelligence measurement, psycho-physiological measurement, psychological assessment, selection and vocational assessment, international comparisons, gender impact, test preparation and coaching, cultural aspects of testing, fairness and bias, adaptive testing, technology and the future of testing. In addition, the exhibition will deal with the challenge: how do we measure characteristics that cannot be measured directly? The exhibition includes some 25 exhibits, most of them interactive. Some of the activities are designed for groups. In addition, the exhibition will include posters, videos, photographs and old test forms. An internet website will be created which can be accessed before and after the exhibition. The exhibition team is already working on the exhibits. It consists of experts in several fields: psychometrics, science, museology and internet design. A public opinion questionnaire has been formulated and data is being gathered and analyzed. We believe that this session will benefit the ITC audience and would be an important addition to the conference program. We conceive the exhibition on testing & measurement as an innovative means of familiarizing the public, youth in particular, with educational and psychological measurement concepts. Dissemination of measurement concepts is one of the organization's main goals. The session offers an open and fruitful discussion with distinguished measurement experts on how to reach the exhibition goals, see below.

Invited symposium 2 15:45-17:15 LT5

The use of ethical principles in testing, - or the lack of it

Chair

Nielsen, Sverre Leonard (The Norwegian Psychological Association, Norway)

Symposium Abstract

This symposium is an attempt to focus on ethical issues in assessment and testing. While quality assurance of tests and testing in general gains more and more attention, the explicit focus on ethics both in regulations and daily use are not so easy to see. Quality assurance of both competence and methods is in itself connected to ethical principles. However, there is still a challenge of raising the ethical awareness among test users.

Paper 1

"Testing ethically: Tensions between principle led ethics and regulatory system"

Lindsay, Geoff (Centre for Educational Development, Appraisal and Research (CEDAR), University of Warwick, UK)*

There has been a large scale interest in the development of ethics by psychologists across the world. The number of countries with an ethical code has increased, stimulated and supported by international initiatives. Within Europe, The European Federation of Psychologists Associations approved a Meta-code of Ethics in 1995, revised in 2005: all member associations are required to have an ethical code compliant and not in conflict with the Meta-code. At a world-wide level, three international associations of psychology (IUPsS, IAAP and IACCP) approved a Universal Declaration of Ethical Principles for Psychologists in 2008 following collaborative, developmental work by an ad hoc group comprising senior psychologists from across the world, a strategy to optimise sensitivity to cultural issues. There have also been capacity building initiatives to support psychological associations in the early stages of development of their code, most recently in South East Europe. Common to many codes developed by national associations is a focus on the use of ethical principles and in some cases supportive material and training in ethical decision making, using the principles and the national code's specifications/standards of behaviour. However, there has also been interest in many countries to attain statutory regulation of the profession, in which case consideration of allegations of unethical conduct may be heard by a separate statutory body with its own code of conduct. In this paper I shall explore the tensions that can arise from the application of each of these two ethical systems

Paper 2

"Ethical guidance when local standards do not exist"

Oakland, Tom (University of Florida, USA)*

Data from two international surveys of test development and use with children and youth, one conducted in 1990 and the other conducted within the last few months, reveal a number of important changes, some of which have important implications for test ethics. These changes are summarized and implications regarding test ethics are described.

Paper 3

"When is assessment a 'Psychological Act'?"

Bartram, Dave (SHL Group Ltd, UK)*

The issue I address is that of deciding when an assessment or some part of an assessment becomes a 'psychological act' and whether 'psychological acts' necessarily require the intervention or input of a practicing psychologist. In considering this issue, we need to consider assessment processes as involving a number of stages and consider how the need for specialist expertise may apply to each stage and under what conditions. I will also consider the degree to which such acts might be carried out remotely or by proxy, and under what conditions. This is an issue that sits at the heart of much of the discussion over the use of the Internet in assessment, especially the issue of remote administration. This has been a topic of much debate over the past few years culminating in the publication of Tippen's (2009) focal paper and accompanying commentaries. I conclude by arguing that ethical assessment depends on the competence of assessors. Professional labels do not guarantee competence and hence we need to relate the nature of the acts carried out in assessment to the competence required by the actors.

Paper 4

"The lack of focus on ethical issues among counsultant within the IO-field"

Hagås, Per Olav (Manpower Professional Executive AS, Norway)*

The issue of ethics is of great importance and affects headhunters, recruitment companies and agencies which let out personnel. Most headhunters and recruiters have a background from sales and management. Few headhunters are psychologists, nor do most of them have adequate competence within objective assessment of human character and/or tools which measure human potentials. Personality tests in particular, but also ability tests are frequently and increasingly used in connection with employments in all categories on any level. The presentation will address the use of test tools and what demands are set to the tests that will affect the careers of the tested ones. Further, what demands that will be profitable expect of the tests and the ones using them? Finally, I will point out the ITC's International guidelines for test use, and propose which parts of these guidelines that should be set as requirements to serious commercial traders. The conclusion is that qualitative demands and ethical norms to a lesser extent have adequate focus, and that this is an issue which is extremely interesting to address.

Invited symposium 3 15:45-17:15 LT6

How to apply personality as a worldwide common concept: Big Five, Big More, or should we? Modeling and usefulness.

Chair

Weekers, Anke (Cito, Netherlands)

Symposium Abstract

The 'Big Five' personality concept is often put forward as a unifying comprehensive framework to describe adult personality worldwide. At the same time, debate goes on regarding four main issues. First, are 'Five' factors enough or are more factors needed to account for the empirical data collected so far. Second, do persons respond to personality items in the way their responses are modeled. More specifically, which factors might give rise to different response strategies. Third, is personality a useful concept in the prediction of organizational behavior for selection purposes, both regarding the incremental validity above other factors and regarding ethical issues. Fourth, how cross culturally valid is any personality framework. The present symposium will discuss the mentioned issues against the background of globalization of test use. Questions like the following then become relevant. Which personality factors are cross culturally valid and which might be specific for which culture? Are there cultural specifics as regards disclosing information on personal behavior? In particular, to what extent is the self report methodology as in personality questionnaires equally applicable in various cultures and which models have to be used to model response processes over various cultures? Are there cultural specifics as regards the predictive validity and ethics of test use in applied psychological practice? Contributors to the symposium will focus on one or more of the above aspects. In a general discussion similarities and differences between cultures will be discussed and consequences will be formulated for both research and practice.

Paper 1

"Impact of measurement invariance on construct correlations, mean differences, and relationships with external correlates: Big Five and RIASEC Measures"

Schmitt, Neal (Michigan State University, USA)*

A relatively large number of cross-cultural studies have investigated the invariance of measures used with various groups and a common finding is that statistically significant differences between groups do exist. In this paper, we evaluate the importance of this lack of invariance on the estimation of structural parameters that relate constructs in these studies. Specifically, the impact of measurement invariance and the provision for partial invariance in confirmatory factor analytic models on factor intercorrelations, latent mean differences, and estimates of relationships with external variables is investigated for measures of two sets of widely assessed constructs: Big Five personality and the six Holland (1985) interests (RIASEC). In comparing models that include provisions for partial invariance with models that do not, the results indicate quite small differences in parameter estimates involving the relationships between factors, one relatively large standardized mean difference in factors between the subgroups compared, and relatively small differences in the regression coefficients when the factors are used to predict external variables. The results provide support for the use of partially invariant models, but there does not seem to be a great deal of difference between structural coefficients when the measurement model does not include separate estimates of subgroup parameters not invariant. Future research should include simulations in which the impact of various factors related to invariance is estimated.

Paper 2

"The Big Five in selection and development assessment: 'Blessing and yoke'

De Fruyt, Filip (Ghent University, Belgium)*

The Big Five has generated a flourishing stream of research leading to a revitalized interest during the past fifteen years in personality assessment in Industrial, Work and Organizational (IWO) psychology. Although this enthusiasm has been welcomed by many, it also has been the subject of considerable criticism (see the debate between Morgeson et al., 2007 versus Ones, Dilchert, Viswesvaran and Judge, 2007; Tett and Christiansen, 2007). Rather than reiterating these arguments, the current contribution highlights new avenues to better align personality research and IWO applications, taking benefit from recent advancements in personality assessment including 'trait-activation theory', 'frame-of- reference research', 'person-centered approaches' and 'a spectrum conceptualization of traits'. Their impact on applied personality assessment will be illustrated using data collected with the Personality for Professionals Inventory (Rolland & De Fruyt, 2009) administered to samples of students and incumbents.

Paper 3

"Response processes in personality measurement"

Weekers, Anke (Cito, Netherlands)*

In the employment context most self-report personality inventories are constructed using classical test theory, factor analysis or dominance IRT models. These models assume dominance response processes, and were originally developed for maximum performance measures (what someone can do), like ability measures. Although personality traits are typical performance constructs (what someone usually does), personality inventories are developed according to the same assumptions as used in maximum performance measurement. However, persons might respond differently to self-report personality inventories, and a different kind of response processes, the unfolding or single-peaked response processes, might be more likely. The usefulness of these response processes will be discussed. An example will be given of an inventory measuring Order that is developed based on single-peaked response processes. The original inventory was developed in the USA (Chernyshenko, Stark, Drasgow, & Roberts, 2007), and translated in Dutch. For this research the translated scale is used, but results will be compared to the results found in the USA.

Paper 4

"Are there Big Cultural differences in Personality?"

Serlie, Alec (Erasmus University Rotterdam / GITP, Netherlands)*
Hiemstra, Annemarie (Erasmus University Rotterdam / GITP, Netherlands)
Van Leeuwen, Rob (GITP, Netherlands)
Bazen, Madelijn (Leiden University, Netherlands)

The results on personality questionnaires play a significant role in selection assessments. Predictive validity of the Big 5 in relation to job performance has been well documented both in America as well as Europe. As personality questionnaires are generally language based, there is a distinct possibility that members of minority groups might either not understand items correctly or misinterpret the meaning of the items. This may in turn bias the results of a questionnaire. Several models of test fairness have been proposed. The most widely accepted is Cleary's model, which is based regression lines. More recently models using Differential Item Functioning (DIF) have been introduced, whereby individual items can be studied in relation to the trait associated with the item. In the present study we set out to study the items of a Five Factor personality questionnaire using various DIF techniques. Our hypothesis was that there would be a difference between (ethnic) minority and (Dutch) majority respondents. In this study the data of a group of 280 test takers (minority: n=119, majority: n= 161) who had completed a FFM personality questionnaire were analyzed. All participant were in their final year of higher (Bachelor or Master) tuition. The results of the study showed that there were significant differences between the two groups on all five factors. On the item level only 11 (5,5%) items showed DIF, of which most could be found in the Neuroticism factor. Items on the Conscientiousness and Openness factors did not show any DIF.

Discussant

"Assessing personality. Does culture matter?"

Van De Vijver, Fons (Tilburg University, The Netherlands & North-West University (Potchefstroom Campus), South Africa)*

Following the four papers put forward at the symposium, their results will be evaluated critically with respect to the following issues: What are defensible empirical generalizations as far as the intercultural (in)variance of a common personality framework is concerned? More specifically, what are the merits of the Big Five factor model in this respect? Do they differ with respect to invariance, or are even additionally other factors necessary to account for intercultural differences? How do the innovative contributions that are put forward with respect to response scale methodology as well as alternative measurement concepts compare with classical big five questionnaire methodology? What are advantages and what might be pitfalls, especially with respect to intercultural differences in response scale use? From a prediction perspective, the incremental validity of more sophisticated scoring methods as well as that of taking situational specifics into account will be discussed. Consequences for advancing theory as well as improving practical utility will be suggested.

Invited symposium 4 13:30-15:00 LT5

The current State of Psychological and Educational Testing and Assessment in Indonesia

Chair

Purwono, Urip (Universitas Padjadjaran, Bandung, Indonesia)

Symposium Abstract

This symposium displays the recent development, current practices, and research in the use and development of psychological and educational testing in Indonesia. The expectation is that participants will have an updated perspective on the landscape of testing in the developing countries such as Indonesia where, in one hand, the number of scholars with adequate background training in psychometrics, test and measurement are limited and, on the other hand, the challenges and work to be done are abundant. The symposium is also aimed at getting thoughtful ideas from the participants. The papers presented in this symposium includes: (1) Mathematics Test Equating under the Graded Response Model; (2) Development and Preliminary Validation of Musical Ability Assessment System; (3) How is IRT applied in personality tests?: A study using Indonesian sample; (4) Improving Students Achievement, Learning Responsibility, and Learning Behavior in Mathematics using Assessment for Learning Model; (5) A closer look at the Indonesia's National Exam Program; and (6) The Uses of Tests in Psychological Practices and Research in Indonesia. Overall, the paper presented will represent (a) the common practices of making assessment instruments available in Indonesian language; (b) the use of test in the common psychological practices and research in Indonesia; (c) current advances in measurement research; and (d) attempts to make a commonly individually administered test available for a large testing program as well as the current trend of collaboration between psychometrician and practitioners as well as researchers in the area beyond psychology and education in Indonesia.

Paper 1

"Improving students achievement, learning responsibility, and learning behavior in mathematics using assessment for learning model"

Mansyur, Mansyur (The State University of Makassar, Indonesia)*

Linking assessment with learning has been a concern in educational assessment for the last two decades. While many attempts have been made to construct assessment that facilitates students' learning, experimental study investigating the effect such assessment program to different aspects of learning was limited. This study investigates whether an assessment for learning program increases (1) student achievement in mathematics, (2) student responsibility in learning mathematics, and (3) student behavior in learning mathematics. A single-group interrupted time-series design (Creswell, 1994) was employed in this study. A total of 10 meetings in-class "Assessment for learning" program was conducted to 244 grade 7 public schools students in Makassar. Assessment instruments consisting of (1) a two stages assignments to assess student's mathematical achievement, (2) self report of learning responsibility, and (3) observation list to be used by teachers were developed. Students' achievement, responsibility, and behavior in each session were estimated employing Samejima's graded response model. Differences of ten data points were analyzed. The results showed that by the end of the program 80.79%, 76.76%, and 71.19% of the students show high achievement, responsibility, and expected behavior consecutively. It is concluded that the assessment for learning program had a positive effect on students' achievement, responsibilities, and behavior in learning mathematics.

Paper 2

"How is IRT applied in personality tests: A study using Indonesian sample"

Halim, Magdalena (Catholic University of Atmajaya, Jakarta, Indonesia)*

Item Response Theory (IRT) is a modern psychometric approach, which provides valuable methods for the evaluation of psychological measurements, including objective personality assessment. Although most published applications of the IRT are used to examine cognitive and ability tests, IRT models are also increasingly being used to study psychometric characteristics of personality tests nowadays. IRT models differ from one another in the number of modeled parameters. Within the family of IRT models, the two-parameter logistic model (Birnbaum, 1968) has been chosen in the present study. The purpose of this study is to show the applicability of IRT analysis in evaluating the psychometric properties of an Indonesian version of the MMPI-2. A total of 1,473 individuals completed a valid MMPI-2; 63.1% were women and age ranged from 17 to 61 (M=23.84, SD=7.38), the rest (36.9%) were men and age ranged from 17 to 66 (M=26.38, SD= 9.29). This sample is part of the Indonesian MMPI-2 normative study. The item parameters for the two-parameter model were estimated with BILOG 3 program (Mislevy & Bock, 1990). The results of this study considerably add to existing knowledge about the psychometric properties of MMPI-2 in Indonesian sample. The two-parameter IRT model fits the data quite well and can be considered as appropriate model. The fact that these results were obtained in an Indonesian sample could be also criticized for being culture specific.

Paper 3

"Development and preliminary validation of Musical Ability Assessment System"

Salim, Djohan (Indonesia Art Institute, Jogyakarta, Indonesia)*
Purwono, Urip (Faculty of Psychology, Universitas Padjadjaran, Bandung, Indonesia)

Musical ability is an important aspect of human quality. From a neurological point of view, individual ability to perform and comprehend musically appears to work independently from other forms of intelligence. This paper reports an attempt to assess individual musical ability using a paper and pencil approach. Defining musical ability as a basic ability to recognize, distinguish, reproduce, and perceive similarity and differences between various sounds, the assessment is carried out by presenting individuals with pre-recorded sounds produced by musical instruments with variation in pitch, timbre, tempo, and dynamics. The individual tasks are to answer questions related to the characteristics of the sound. Standardized is done by pre-recording the sounds in a CD. A total of 200 participants were recruited for the study. Partial EEG activities were also recorded for validation purposes. Even though, in term of internal consistency, the estimated reliability was at the lower range, the psychometric properties of the assessment system show promising features. Preliminary validation also shows consistencies with neurological theory, suggesting that the idea underpinning the assessment can be developed further, particularly to facilitate research in the area of psychology and music.

Paper 4

"A closer look at the National Exam Program in Indonesia"

Mardhapi, Djemari (Indonesia National Education Standard, Indonesia)*

Since its conception back in early '60, Indonesia National Exam has been a concern to parents, educators, and policy makers. As a results of the controversies, political agenda, and advances in test theory, the format of Indonesia National Exam has been undergone several changes. This paper presents the historical, technical, delivery, and reporting aspects of the National Exam in Indonesia and the changes in its format from one period to another. Information is gathered from direct observation, reports, printed documents, publications and governmental official archives related to the design and implementation of the exam. Results of the analysis to the item level national exam data were also examined. It is asserted that, from psychometric perspectives, the current national exam has sound technical aspects. The IRT one parameter logistic models was appropriate for the data. Arising problem is hypothesized as more closely related to the non technical aspects of the exam, particularly its administration and the public misperception concerning the exam.

Paper 5

"The uses of tests in psychological practices and research in Indonesia"

Suhapti, Retno (Indonesian Psychological Association, Indonesia)*

The history of psychology in Indonesia is tied with the use of psychological testing for personnel selection, placement and classification. This tradition continues until recently, and this paper identifies issues around psychological testing and its uses in Indonesia. Data from this research was obtained from a survey administered to professional psychologist in Indonesia inquiring activities frequently engaged in their services and their uses of test materials. In addition, information was also gathered from three separate large group discussions taking places in three difference provinces. A focused group discussion was also conducted to revalidate the information. Results suggested the questionable practices related to the development and use of test materials which stemmed from inadequate background training in testing and measurement coupled with lack of guidelines and regulation concerning test uses in the country. Another finding suggested that the most frequently used psychological testing instruments were transported from either the US or Europe. However, proper adaptation of these instruments and empirical investigation of its equivalences with the parent language culture of the test is rarely conducted posing a validity concern to their uses in Indonesia. These findings are then discussed in the light of the history psychological education in Indonesia, particularly those related to testing and measurement. Recommendations for future intervention, particularly those related to the international testing community, are presented.

Invited symposium 5 15:15-16:45 LT1

Future directions of testing in the United States

Chair

Hambleton, Ronald K. (University of Massachusetts, USA)

Symposium Abstract

The importance of test uses continues to grow. In the USA today, students from the 3rd grade to high school are administered achievement tests, for a total of about 80 million tests per year. Add several times 80 million tests to account for the diagnostic tests that are being administered to support the assessment of student progress, and it is clear that the growth of testing in the schools has been substantial. Also, the number of admissions tests and credentialing exams continues to grow. At the same time, because of the importance of these tests, and the desire to improve test score validity, technical advances in many directions are occurring. In this symposium, the presenters will focus on three important directions in the USA: First, we will consider the impact of cognitive psychology on testing. The impact has been discussed in the measurement literature for 30 years but now that impact is being seen, and substantial amounts of research are underway. Second, the impact of technology has been increasing, and now it is likely to fundamentally change our approaches to what is measured, as well as to test design and test administration. Advances in technology will be the focus of the second presentation. Finally, nothing is more important, ultimately, than the production of valid test scores, that can be reported on meaningful score scales, and in ways that they are understood and used correctly by practitioners. Score reporting in the future will be the focus of the final presentation.

Paper 1

"Using advances in learning and cognition to improve assessment in the 21st Century"

Huff, Kristen (The College Board, USA)*

Progressive approaches to curriculum and instruction are strongly influenced by what we are learning about how students develop deep conceptual knowledge, critical thinking skills and the propensity to learn more advanced topics. As such, contemporary assessment design needs to be aligned with these more complex targets of learning that are valued in the classroom versus those typically measured on large-scale standardized tests such as factual recall and "plug and chug" procedures. Assessments of the future will, for example, evaluate students on the quality of the questions they have about a new topic rather than measure their mastery of content, and provide feedback that helps teachers move students along the stated learning path. In addition, strong validity arguments for contemporary assessments will include a rationale for how items are designed to explicitly target the knowledge and cognitive processes of interest. This presentation will provide an overview of progress in these areas (e.g., evidence-centered assessment design), as well as areas in need of more research (e.g., dynamic assessments of transfer).

Paper 2

"Impact of technology on testing practices"
Mills, Craig (American Institute of Certified Public Accountants, USA)*

Technology is transforming business and business processes at an ever accelerating rate. Concurrently, emerging economies are providing inexpensive access to labor, both skilled and unskilled. Widespread access to the internet and social media has changed how professionals, students, youth, and adults work, play, and interact with one another. Inevitably, these trends will impact measurement practice as well. We can anticipate that current batch processing for item development, test administration, and statistical analyses will be increasingly replaced by continuous process flows. These process flows will also become increasingly automated and decentralized. Additionally, traditional testing formats may well become less valid as success in education and work become more dependent on collaboration and use of disparate data resources. Tests will need to change in order to assess entry-level skills or educational promise as the skills needed for success change. This presentation will suggest changes in testing processes and tests themselves that are responsive to the changing needs of test sponsors and test takers. It will draw from lessons learned in other industries and apply them to test development, administration, and analysis.

Paper 3

"Next generation of Latent Variable Models: New opportunities and challenges for testing"
Zumbo, Bruno D. (University of British Columbia, Canada)*

In recent years, many exciting developments have taken place in latent variable modeling, but perhaps none more so than the development of methods that (a) explore and account for unobserved population heterogeneity, or mixtures of unobserved groups, (b) mixtures of continuous and categorical latent variables, (c) multilevel models for complex measurement data, and (d) approaches for dealing with categorical item response data. This next generation of latent variable models shows tremendous potential for improving testing practice and research by expanding the types of modeling being done and hence aligning more closely with the complex nature of contemporary tests and testing contexts. I will present an overview of this next generation of advanced methodology and describe how it may inform testing from day-to-day practices such as dimensionality assessment to more theoretical work that focuses on evaluating explanatory models of test data for the purposes of supporting validity research and practice.

Paper 4

"Improving the score scales and reports we use in communicating test results"
Hambleton, Ronald K. (University of Massachusetts, USA)*

Testing practices in education and psychology have advanced considerably in recent years through the introduction of item response theory models, generalizability theory, automated test assembly, and new test designs such as computer-adaptive testing. At the same time, methods for reporting test scores and diagnostic information to candidates, the culmination of the testing process, remain largely understudied and undervalued as a problem in educational and psychological assessment. This is most unfortunate too because of the large amount of evidence suggesting that candidates and other score users such as teachers, policy-makers, psychologists, and the media, are often confused by the meaning of test scores resulting in misinterpretations, and candidates are often disappointed by the limited amount of diagnostic information they receive from hours of testing. The goals of this presentation include: (1) highlighting several promising ideas (e.g., bench-marking and item mapping) for increasing the clarity and meaning of score reports, and (2) offering some steps, based on emerging research, for preparing score reports and score scales. Examples of good practices will be presented and come from our work with several national and state testing programs, and credentialing agencies in the United States.

Invited symposium 6 15:15-16:45 LT5

Recent developments in the assessment of emotional intelligence

Chair
Fontaine, Johnny (Ghent University, Belgium)

Symposium Abstract

Since its inception the construct of emotional intelligence (EI) has attracted a lot of attention from both practice and science. While being very popular in practice, the construct has been severely criticized in the scientific arena. The EI trait approach, which uses self-reports, is reproached to lack discriminant validity. The EI ability approach, which uses maximum performance tests, is criticized for the content of its items and for its scoring. In this symposium four recent developments that take these criticisms to heart are being presented. In the first contribution a set of four new assessment instruments is proposed. They form a multi-method approach in which EI is assessed by self-reports, other-reports, ability items that can be scored as correct or incorrect, and coding of interviews. The second contribution proposes a whole new approach to the assessment of emotional abilities by using multimedia applications that greatly enhance the ecological validity of the assessment procedures. In the third contribution a plea is made to embed the assessment of emotional intelligence in current emotion theory. Based on a componential emotion perspective theoretically-grounded and empirically-based scoring keys are proposed. Finally, very critical evidence is presented on the context-free assessment of emotional expression, which often forms a key part of emotional intelligence testing. As the four contributions differ with respect to their evaluation of the EI construct, the symposium will end with a discussion on how these recent developments overcome or just confirm earlier criticisms.

Paper 1

"Alternative methods assessing the emotional intelligence of chinese respondents"

Wong, Chi-Sum (The Chinese University of Hong Kong, Hong Kong SAR, China)*
Peng, Kelly (Hong Kong Shue Yan University, Hong Kong SAR, China)
Huang, Emily (Hong Kong Baptist University, Hong Kong SAR, China)

There are severe criticisms on the construct and the measures of emotional intelligence (EI) in late 1990s (Davies, Stankov, & Roberts, 1998). In response to these criticisms, we have been developing four different methods to assess Chinese EI in the past 12 years according to the ability model of EI, i.e., EI is defined as the ability to deal with emotions rather than personality. The first method is the development of a self-report measure. The development process and evidence of reliability and criterion-related validity is reported in Wong and Law (2002). The second method uses other-rated items. That is, raters are evaluating people whom they are familiar with. Evidence of reliability and criterion-related validity for this method is reported in Law, Wong and Song (2004). The third method is to develop test items that have correct versus incorrect options. The development process and evidence for reliability and criterion-related validity is reported in Wong, Law and Wong, 2004 and Wong, Wong and Law, 2007. The final method is to assess Chinese EI level by specific questions used in selection interview. We have gathered reliability and validity evidence for some carefully developed situational interview and behavioral interview questions. Details of the developmental process, and the pros and cons of the above four methods are discussed.

Paper 2

"Multimedia assessment of emotional abilities: Research and development"

Roberts, Richard (Educational Testing Service, USA)*
Schulze, Ralf (University of Wuppertal, Wuppertal, Germany)
Minsky, Jennifer (University of Wuppertal, Wuppertal, Germany)
MacCann, Carolyn (University of Sydney, NSW, Australia)

Research examining emotional abilities (EA) is in the ascendancy, with several target articles in influential journals such as the American Psychologist and the Annual Review of Psychology. However, limitations in extant assessments and the need for alternative measurement approaches are apparent. We discuss the development of two multimedia tests: (1) A situational judgment test (where participants rate a scenario for emotional relevance and salience) and (2) a principal-agent paradigm (where event-emotion contingencies in others have to be perceived and memorized and emotion-behavior contingencies inferred from observed behavior to predict future behavior). In two studies (N=857) these EA assessments were administered to community college and university students across the USA. Study 1 evaluates the psychometric properties, including tests for measurement invariance and examination of subgroup differences (e.g., ethnic groups). We also present the new tests' relationships with emotions measures (e.g., the Mayer-Salovey-Caruso Emotional Intelligence Test), outcome measures (such as GPA and coping with stress), personality (as assessed by the Big Five), and five broad cognitive ability factors (i.e., fluid, crystallized, fluency, spatial, and quantitative ability). Study 2 examines test-retest reliability, along with relationships that the measures share with positive affect, as assessed by the Day Reconstruction Method. Overall, findings suggest that multimedia assessments of EA are reliable and share meaningful relations with (a) crystallized intelligence, (b) emotions measures, and (c) valued outcome variables (e.g., coping with stress). We conclude with a discussion of some limitations and future research that aims to address identified problem areas and extend the multimedia approach.

Paper 3

"Constructing scoring keys for the assessment of emotion knowledge"

Fontaine, Johnny (Ghent University, Belgium)*
Scherer, Klaus (University of Geneva, Switzerland)

A major issue for the viability of the emotional intelligence construct is the scoring key of the ability items. How can one determine the correctness of an answer to a question about emotions? The three main approaches, namely consensus scoring, expert scoring, and target scoring, have been severely criticized. They would not be accepted as valid ways to identify correct answers of classical intelligence items. In the present paper an alternative approach is proposed for one aspect of the EI construct, namely for emotional knowledge. The approach is based on the GRID instrument, which consists of 142 features that operationalize six emotion components (appraisals, expression, subjective experience, bodily reactions, action tendencies, regulation) and 24 emotion terms (for the assessment of meaning) or daily emotional episodes (for the assessment of experiences). It probes the salience of each emotion feature for each emotion word or for each daily episode. In a first study in Belgium, Switzerland, and the UK (Fontaine, Scherer, Roesch, & Ellsworth, 2007), a principal component analysis revealed a robust overall four-factorial structure (pleasantness, potency, arousal, and unpredictability) for the meaning of emotion words. This structure was confirmed in a recent, large cross-cultural study in 30 linguistic/cultural groups from five continents and in a large-scale emotion episode study in Belgium. It will be discussed how the results from the GRID studies can be used to derive theoretically-grounded and empirically-based scoring keys for emotional knowledge items.

Paper 4

"Emotion judgments are relative: Implications for assessing emotional intelligence"

Yik, Michelle (The Hong Kong University of Science and Technology, Hong Kong SAR, China)*
Zeng, Kevin (The Hong Kong University of Science and Technology, Hong Kong SAR, China)

Judging others' emotions is central to daily social interactions and is the basis upon which teachers, parents, lovers, and friends behave. The ability to make accurate emotion judgments was used as a benchmark for distinguishing people diagnosed with schizophrenia or ADHD from controls (Kerr & Neale, 1993; Rapport, Friedman, Tzelepis, & Voorhis, 2002) and for assessing individual differences in emotional intelligence (Roberts, Zeidner & Matthews, 2001). Pertinent to this everyday wisdom

are the assumptions that we automatically express our emotions via verbal or nonverbal behaviors and that observers, with minimal efforts, are capable of efficiently decoding the behaviors and correctly judging the emotions expressed by others. In the present study, we examined the mechanism underlying the process of judging others' emotions from the anchoring and adjustment perspective. We showed that judgments of emotions communicated in emotion scripts were influenced by the context for judgment (viz. an anchor). Our results challenge the practice of using emotion judgments as a yardstick to measure emotional intelligence.

Discussant

Grégoire Jacques (Université Catholique de Louvain, Belgium)*

Invited symposium 7 15:15-16:45 LT4

Recent developments of CBT in Japan

Chair

Shigemasu, Kazuo (Teikyo University, Japan)

Symposium Abstract

Computer Based Testing has been finally getting popular in Japan. In this symposium, we will introduce some unique efforts both in terms of theory and practice to promote CBT in Japan.

Paper 1

"Implementing multidimensional item response models for routine computer based testing"

Muraki, Eiji (Tohoku University, Japan)*

Compensatory and non-compensatory multidimensional item response theory (MIRT) models have been constructed and their parameters are estimated by the marginal maximum likelihood (MML) method. Any cognitive tests seem to be hardly unidimensional because their cognitive tasks require various combinations of complex mental functions. However, it is quite difficult to use routinely MIRT models to the standardized testing situations because the MIRT models are built to aim essentially at capturing the interactions between test items and subjects' complex cognitive performances and those interactions can be thought to qualitatively differ among subgroups of subjects, such as their gender and ages. The routine implementation of the MML method is also causing problems because the estimation method needs multiple integrations and their complexity increases exponentially as the number of dimensions is added. In this presentation, the MCMC method is derived to estimate the parameter values of the MIRT models and suggest a reasonable procedure which can be implemented routinely for standardized computer-based testing applications.

Paper 2

"Introduction to the Common Achievement Test System for entering clinical clerkship in Japanese medical schools"

Mayekawa, Shinichi (Tokyo Institute of Techonology, Japan)*

Common Achievement Test was introduced in 2005 in order to assess the students' mastery of the core curriculum before entering clinical clerkship in Japan. The test was developed using IRT models and administered through the network of computers.

Paper 3

"Challenges in developing and operating CBTs in Japan"

Nogami, Yasuko (The Japan Institute for Educational Measurement, Inc., Japan)*
Kobayashi, Natsuko (The Japan Institute for Educational Measurement, Inc., Japan)
Hayashi, Norio (The Japan Institute for Educational Measurement, Inc., Japan)

It has been about eight years since the Japan Institute for Educational Measurement (JIEM) released the CASEC, a computerized adaptive testing system to measure proficiency of English as foreign language. Through Internet, examinees can take the test at any time and any place, and they receive feedback immediately upon completion of the test. The number of examinees has been steadily increasing. In 2009, the test was taken more than 110,000 times. The examinees range widely in background—from junior-high-school students to university graduates and adults in the workforce. The results of the test are used for different purposes and in different contexts; placement in schools, monitoring educational achievements, and so on. The JIEM also has another type of computerized test called CASEC-G. Examinees are required to translate Japanese sentences into English ones, and their writing skills are evaluated. After taking the test, examinees receive some advice on how to improve their performance and they can brush up their writing skills with a tutorial system called CASEC-GTS. In addition, the JIEM will release a new computerized test to assess reading skills of English in April 2010. We would like to introduce these computer based tests developed and operated by the JIEM. We will illustrate some difficulties we encountered in the process, and our efforts to solve them.

Paper 4

"*A nationwide listening comprehension test using personal IC players"*

Otsu, Tatsuo (The National Center for University Entrance Examinations, Japan)*
Uchida, Teruhisa (The National Center for University Entrance Examinations, Japan)
Ito, Kei (The National Center for University Entrance Examinations, Japan)

A Japanese scholastic standard nationwide examination, called the National Center Test (NCT) is conducted by NCUEE in January every year. All national and local public universities as well as part of private universities make use of NCT. Usually, each university administers its own tests in February of March. Currently, the use of NCT by private universities is supplementary. Usually they assign a small portion of the admissions for NCT applicants, and the rest are assigned to applicants for their own examination. There were more than a half million applicants participated in NCT in 2009. NCT is designed to assess the basic scholastic achievements which applicants have attained in upper secondary high school. NCT 2009 provides tests in 28 subjects in six areas, Japanese language (including Japanese and Chinese classics), geography and history, civics, mathematics, sciences, and foreign languages. Every applicant is not required to take all the six subject areas, but each university designated the subject areas or subjects at its discretion. English test of NCT contains listening comprehension test items in addition to usual paper and pencil test items. The NUCEE conducted the listening test in 2006 at the first time. The listening comprehension test of NCT consisted of 25 short questions, and took 30 minutes. We will introduce our operation of the test, contents of the test items, and its influences on the university admissions in Japan.

Discussant

Zhang, Houcan (Beijing Normal University, China)*

Wednesday, 21st July 2010

Invited symposium 8 08:30-10:00 LT5

The Wechsler intelligence scales cross the globe: Measurement variance and invariance.

Chair

Weiss, Lawrence G. (Pearson, USA)

Symposium Abstract

This symposium is about the measurement invariance and variance of the Wechsler Intelligence Scales. Up to date, most research about the measurement invariance and variance of the Wechsler intelligence scales was conducted using data from the west globe. With the publication of WISC4 and WAIS4 in Taiwan, Hong Kong, and Mainland China, we will report our recent findings about the measurement variance and invariance of the Wechsler Scales using the Asian data. Four papers related to measurement invariance, language effect, Flynn effect, and validity evidence of the Wechsler scales will be reported. The theoretical, practical, and clinical meanings of the results will be discussed within the context of previous studies using the data from the west globe.

Paper 1

"The measurement invariance of WISC4 Hong Kong, Macau, Taiwan, and Mainland China."

Chen, Hsin-Yi (Pearson, USA)*
Weiss, Lawrence G. (Pearson, USA)
Li, Yuqiu (Beijing Normal University at Zhuhai, China)

The purpose of this study was to test measurement invariance of the WISC-IV factorial structure between four regions: Hong Kong, Macau Taiwan, and Mainland China. The structure reported in the US WISC-IV manual (Wechsler, 2003) was used as the hypothesized baseline model. Then, multi-sample analyses were conducted with constraints embedded in a stepwise manner. We tested for invariance on four levels of nested models. Each level had more constraints than the previous one (Meredith, 1993). The first and weakest level was configural invariance. It assumed the overall factor pattern was the same between regions. The second level was testing for weak factorial invariance, also called metric invariance. This model required the magnitude of the factor loadings to be the same between regions. The third stage tested unique variance invariance, which examines whether the test measures the same construct with similar accuracy. Finally, in the most restricted model, the factor covariances were all constrained to be equal across genders. All factor models were tested using covariance matrices. Maximum likelihood was the estimation method chosen because of its robustness and sensitivity to incorrectly specified models. During each step of the analyses, the chi square difference (Δx2) was tested between nested models and suggestions regarding partial measurement invariance were carefully considered and followed. If inadequate fit was detected, fit in the model was improved by including additional parameters identified by the modification index (MI) provided by LISREL. Re-parameterization was examined carefully for meaningfulness.

Paper 2

"Language effects on the performance of WISC4 subtests: Evidence from the U.S., Hong Kong, Macau, Taiwan, and China."

Li, Yuqiu (Beijing Normal University at Zhuhai, China)*
Zhu, Jianjun (Pearson, USA)
Chen, Hsin-Yi (Pearson, USA)

The current study evaluates the language effects on the performance of WISC4 subtests. First, we hope to replicate the results reported by Chen and Zhu (2004) using the WISC4 data from multiple samples. Second, we want to evaluate the language effects on children's performance of the Digit Span subtest. Because the numbers (digits) used in Chinese language are short and simple, they are much easier to memorize and pronounce. As a result, children from Hong Kong, Macau, Taiwan, and Mainland China should show higher digit span forward and backward scores than U.S. peers. Samples matched on parent education, age, and sex were drawn from U.S., Hong Kong, Macau, Taiwan, and Mainland China normative samples. Next, children's performance on the WISC4 Coding, Symbol Search, Digit Span, Matrix Reasoning, Block Design, and Picture Concept subtests were compared across the five matched samples.Preliminary results confirmed the previous finding by Chen and Zhu (2004). Children from Hong Kong, Macau, Taiwan, and Mainland China did significantly better on Coding and Symbol Search subtests than American peers. In addition, children from Hong Kong, Macau, Taiwan, and Mainland China also did significantly better on Digit Span Forward and Backward subtests. On average, U.S. children scored 4.5-4.9 points lower on Digit Span Forward and 0.5-2.2 points lower on Digit Span Backward. The theoretical, practical, and clinical implications of these results on cognitive research, test development, and clinical practice will be discussed.

Paper 3

"Cross-culture comparison of Flynn effect on Wechsler Intelligence Scales"

Zhang, Houcan (Beijing Normal University, China)*
Zhu, Jianjun (Pearson, USA)
Chen, Haipin (Beijing Normal University, China)
Chan, Yat (Educational Psychology Service, China)

For more than two decades, research has shown consistent support for the Flynn effect on Wechsler intelligence scales. However, due to a shortage of data, there are very few studies specifically evaluating the Flynn effect on the Wechsler intelligence scales using Asian samples. The current study will evaluate the Flynn effect on Wechsler intelligence scales using data from Taiwan, Hong Kong, and Mainland China. Data from the four validity studies will be used to evaluate the Flynn effect. Composite scores will be used in the data analysis. If possible, data from the four validity studies will be pooled to increase the statistical power of the current study. The current study will be focusing on the following research questions: (1) Are the Flynn effects observed from these four studies consistent with the expectation set forth by Flynn (1984, 1987), i.e., about a 0.3 increase in FSIQ points per year? (2) Are the Flynn effects observed in the current study invariant from those reported in U.S. edition of the Wechsler manuals? (3) Are the Flynn effects observed in the current study invariant across the four samples? Are there any age trend? (4) Are the Flynn effects observed in the current study invariant across all ability levels? The theoretical, practical, and clinical meanings of the results will be discussed within the context of previous studies that evaluate the Flynn effect on Wechsler scales.

Paper 4

"Evidence of reliability and validity of WAIS4 China."

Zou, Yizhuang (Beijing Huilongguan Hospital, China)*
Wang, Jian (Beijing Huilongguan Hospital, China)

The Chinese adaptation of the U.S. WAIS4 is currently in progress, and the standardization of the instrument will be finished by early 2010. As part of the standardization, the following Chinese samples will be collected: a normative sample, a test-retest sample, an inter-scorers reliability sample, a paper-penciled and digital administration equating sample, and a couple of clinical samples. The current presentation will focus on the flowing psychometric properties of the Chinese WAIS4: (1) Representativeness of the normative sample; (2) Evidence of reliability, such as internal consistency reliability, test-retest stability, and inter-scorer agreement; (3) Evidence of validity, such as exploratory and confirmatory factor analyses, inter-subtest correlations, correlation between the previous and current edition, and the equivalency of paper-pencil and digital administration; and (4) Initial evidence of clinical validity based on a sample of individuals diagnosed with schizophrenia and a sample of individuals diagnosed with mental retardation. The consistency between the psychometric properties of the Chinese WAIS4 and the U.S. edition will also be discussed.

Discussant

"The past, current, and the future of the research on the Wechsler intelligence scales"

Grégoire, Jacques (Université Catholique de Louvain, Belgium)*

The discussion will be focus on the following: (1) A brief review about the previous cross cultural research on the Wechsler intelligence scales; (2) The theoretical, practical, and clinical implications of the four papers presented at the symposium; (3) the direction of the future cross-culture research on the Wechsler scales.

Invited symposium 9 08:30-10:00 LT6

A new generation of DIF studies

Chair

Elosua, Paula (University of the Basque Country, Spain)
Hambleton, Ronald K. (University of Massachusetts, USA)

Symposium Abstract

Numerous DIF studies have been published in specialized and applied psychometric journals during the last two decades. In addition to the development of statistical procedures for detecting differential item functioning that are highly efficient in spotting problematic items, the research on DIF to date has also focused on applications of DIF analyses in a range of testing contexts. All of this work is critical because of the extent to which DIF analyses are a fundamental part of item analysis. However, it is important to note that any analysis of differential item performance should not be narrowly focused on the detection of DIF: once DIF is detected, the task turns to understanding it, the study of effects of item type on examinee performance, or the study of the practical consequences. It is this idea of extending DIF studies with new methods and approaches that forms the basis of this proposed symposium: A new generation of DIF studies. The new perspective involves multilevel latent models, mixed models, consequences and new robust procedures for the detection of DIF. The symposium consists of four presentations given by researchers from four countries. The first study illustrates a new approach to detecting DIF based on using robust statistics ; the second one uses a simulation to evaluate the effects of factorial partial invariance on group comparisons ; the third and fourth presentations incorporate mixture models to evaluate the presence of latent classes and novel applications of multilevel IRT .

Paper 1

"Robust anchoring and posterior anchoring as procedures for DIF and measurement equivalence"

de Boeck, Paul A. L. (University of Amsterdam, Netherlands)*

An important issue in the process of identifying DIF and also in the process of obtaining measurement equivalence is the choice of anchor items. The basis for this choice is commonly either prior knowledge or iterative purification based on the data. Two alternatives are presented here: (1) robust anchoring, using tools from robust statistics, and (2) posterior anchoring, based on posterior DIF probabilities of the items. The robust approach can be implemented in a parametric way, for example with a robust version of the Raju distance, or in a nonparametric way, for example with marginal proportions correct. The posterior approach requires a mixture model for the items, with a DIF class and a non-DIF class of items. These two alternatives do not require prior knowledge and neither do they make use of iterative purification. They both rely on a one-step statistical procedure. Simulation studies show that their performance is excellent. Apart from their practical use in dealing with DIF and obtaining measurement equivalence, they are also novel IRT approaches in a more fundamental statistical sense.

Paper 2

"The effect of Partial Factorial Invariance on group comparisons"

Elosua, Paula (University of the Basque Country, Spain)*
Zumbo, Bruno D. (University of British Columbia, Canada)

Factorial invariance studies examine the equivalence among factorial structures across groups. Conclusions about partial factorial invariance mean that some of the model parameters (loadings, thresholds, error variances) are different for groups. It is difficult, however, for a researcher to quantify the effects (i.e., impact) of this lack of invariance on subsequent statistical decisions based on group mean comparisons or coefficient alpha comparisons across groups.

Paper 3

"Latent variable mixture modeling as a method to examine sample heterogeneity, and the related problem of DIF"

Zumbo, Bruno D. (University of British Columbia, Canada)*
Sawatzky, Richard G. (Trinity Western University, Canada)
Ratner, Pamela A. (University of British Columbia, Canada)
Kopec, Jacek A. (University of British Columbia, Canada)

We will present an overview of a program of research that applies latent variable mixture modeling (LVMM) to examine the extent to which a sample is homogeneous with respect to a specified statistical model for ordered categorical item responses. Along the way we will evaluate the implications of sample heterogeneity with respect to the latent variable scores, and identify potential sources of sample heterogeneity. As has been shown in the literature, LVMM can be used in conjunction with IRT (i.e., an IRT mixture model) to examine sample heterogeneity, and the related problem of DIF, when relevant group differences are not assumed a priori (Cohen & Bolt, 2005; De Ayala et al., 2002; Mislevy, Levy, Kroopnick, & Rutstein, 2008; Rost, 1990; Samuelsen, 2008; Vermunt, 2001). Our aims are: (a) to share the lessons we have learned about LVMM, its implementation and limitations, and (b) demonstrate how looking at the typically DIF situation from this vantage point allows us to investigate whether there are other variables than the usual manifest variable in DIF studies (such as gender, age, or nationality), or interactions among variables, that distinguish homogeneous groups. Our focus will be typical psychosocial measures such as emotional wellbeing and physical functioning, and the data complexities they present.

Paper 4

"Applications of multilevel IRT models to investigate item type effects"

Zenisky, April L. (University of Massachusetts, USA)
Elosua, Paula (University of the Basque Country, Spain)
Zumbo, Bruno D. (University of British Columbia, Canada)*

This presentation focus on a new multilevel IRT model and on its application to study item type effects which can affect the performance across groups. A multilevel IRT model developed for group-level diagnosis was applied to study data from high school end-of-course examinations. Variability in item difficulty across ethnic groups was investigated in relation to item features associated with content and cognitive process categories. Random effects were attached to each feature type at the group level, and their variability studied across groups. The estimated feature effects were shown to provide a basis for examining cross-ethnic differences for individual features as well as cross-feature differences within individual ethnic groups, as this may be useful for diagnostic purposes. The model was fitted using Markov Chain Monte Carlo procedure by R software.

Invited symposium 10 10:00-11:30 LT5

Development of psychological test in Mainland China

Chair
Zhang, Jianxin (Institute of Psychology, Chinese Academy of Sciences, China)

Symposium Abstract
Four speakers will present separately their papers on the application of psychological measurements such as MMPI and CMHI, and on use of the new techniques such as IAT in developing tests, and on Ethical Code of psychological tests endorsed in Chinese mainland.

Paper 1

"Theory and method of psychological and educational measurement being widely applied in Chinese mainland"
Zhang, Minqiang (Southern China Normal University, China)*

In Chinese mainland, the candidates of any test are numerous because of a huge population. The number of university entrance exam takers has reached 10 million, and the number of candidates participating in the entrance exams for postgraduate schools has increased to 1.2-1.5 million. Moreover, there are a large number of candidates in other examinations, such as the judicial examination, the qualified doctor practitioner examination, the accountant qualification test, the civil service examination, and even in tests for foreign candidates, such as HSK. The increasing application of test and the improvement of demand for test organization has strongly pushed the theoretical and application research of psychological and educational measurement in Chinese Mainland. A psychometric committee has been established under Chinese Psychological Society; under the Chinese Society of Education, there is a branch for educational measurement and statistics, with approximately 1,000 members. These professionals active in all kinds of field are prompting the theoretical and application studies in psychometrics, and great success has been achieved in theory and application of CTT, GT and IRT. Plenty of scales with conformance to psychometric rules have been widely used for different kinds of population, which play an important part in improving the mental health of Chinese people as well as preventing the mental disease.

Paper 2

"The 2008 revision of the Chinese Code of Ethical Use of Psychological Tests"
Gan, Yiqun (Beijing University, China)*
Che, Hongsheng (Beijing Normal University, China)

In response to the rapid increase of application and abuse of psychological tests in China, the Psychometrics Division of Chinese Psychology Society (CPS) made major revisions to the Chinese Code of Ethical use of Psychological Tests in 2008. Comparing to the earlier version in 1992, the rules were reorganized to define more specifically the responsibilities of test users and the rights of test takers. New items concerning the test users' qualification and the validation of instruments were added. The test users are recommended to use only those psychological tests approved or registered by the CPS. In addition, a number of points relevant to the respect for test takers' rights and privacy were stated more explicitly. The current code provides the psychological test users in China clear guidelines for ethical decision making in their work.

Paper 3

"The arena of mental health measurements in Mainland China: From SCL-90 to CMHI"
Chen, Zhiyan (Institute of Psychology, Chinese Academy of Sciences, China)
Wu, Zhenyun (Institute of Psychology, Chinese Academy of Sciences, China)
Huang, Zheng (Institute of Psychology, Chinese Academy of Sciences, China)*
Guo, Fei (Institute of Psychology, Chinese Academy of Sciences, China)

The translation and normalization of many foreign mental health measurements has been done since 1980's. In the past 30 years, SCL-90 has been the most used mental health measurement in college and hospital settings. Other often used instruments in college settings and in hospital settings differed. The former were UPI, EPQ, and 16PF, the later were SAS, SDS, HAD, HAMD, BDI, etc. As the most used mental health measurement in mainland China, SCL-90 has been criticized for its improper application in community sample, inability to identify "negative symptoms", and so on. To provide an instrument more applicable in non-patient sample, Chinese Mental Health Inventory (CMHI) was developed to measure individual's level of mental health with psychological concepts rather than psychiatric symptoms. CMHI has five dimensions, including emotion experience, self-evaluation, interpersonal capacity, cognitive efficacy and adaptiveness. Aside from the attempt to assist diagnosis of several mental disorders, follow-up mental health service system after measurement was also provided for CMHI.

Paper 4

"The clinical application of the MMPI in Mainland China"
Wang, Li (Institute of Psychology, Chinese Academy of Sciences, China)*
Zhang, Jianxin (Institute of Psychology, Chinese Academy of Sciences, China)

The Minnesota Multiphasic Personality Inventory (MMPI) was first introduced into mainland China in the early 1980s. As an objective personality test with sound psychometrical properties, the MMPI rapidly became one of the most popular assessment instruments in clinical setting in mainland China. It was widely used as a screening or an aided diagnosis tool in variety of populations, and has been demonstrated to have useful clinical applications. However, given that the short history of the MMPI in mainland China, its clinical application is mainly limited to basic clinical scales and a few content scales. Therefore, more studies should be conducted to further validate new developed content scales, additional scales, and special scales for promoting their clinical application in Chinese population.

Discussant

Zhang, Houcan (Beijing Normal University, China)*

Comments on the above four speakers' presentations in particular, and on Chinese psychological tests in general will be provided.

Assessment models for monitoring learning

Chair

Hambleton, Ronald K. (University of Massachusetts, USA)

Symposium Abstract

The symposium will discuss on the models for monitoring teaching and learning in three countries: Denmark, Hong Kong and New Zealand. The three systems will be reviewed and discussed in terms of their influences on learning and teacher autonomy, the stakes associated with assessments, the types of assessments used, the levels of aggregation of data from these assessments, and how data are used.

Paper 1

"National tests in Denmark – CAT as a pedagogic tool"

Wandall, Jakob (Danish Ministry of Education, Denmark)*

Testing and test results can be used in different ways. They can be used for regulation and control, but they can also be a pedagogic tool for assessment of student proficiency in order to target teaching, improve learning and facilitate local pedagogical leadership. To serve these purposes tests have to be low stake. In Denmark, to ensure this, test results are made strictly confidential by law. The only test results that are made public are the overall national results. Because of the test design (Rasch-model), results are directly comparable, which gives an enormous potential for monitoring added value and developing new ways of using test results in a pedagogical context. The presentation gives the background and status for the development of the Danish national tests, describes what is special about these tests (IT-based, 3 tests in 1, adaptive, etc.), how the national test are carried out and what is tested. Futhermore, it is described who are allowed to know the results, what kind of response is given to the pupil, the parents, the teacher, the headmaster and the municipality and how the results can be used by the teacher and headmaster.

Paper 2

"Alternatives to external standardized assessments: Hong Kong example"

Hamp-Lyons, Liz (University of Hong Kong, Hong Kong SAR, China)*

In this presentation I will 1) describe the school-based assessment system that has been introduced across Hong Kong secondary education to assess the English speaking skills of all students; 2) describe how this classroom assessment data is used to report student level data for educational planning and region-wide accountability; 3) discuss how and to what extent this school-based assessment supports learning in the classroom and contributes to teacher professional development.

Paper 3

"Assessment models for monitoring learning: New Zealand"

Hattie, John (The University of Auckland, New Zealand)*

New Zealand has a recent history of self-managed schools with many freedoms to make decisions about teaching and assessment. There are many options for them to choose. The session outlines the options available in an on-line assessment package (asTTle) which includes Teacher customised, comprehensive, computer adaptive, interview, and attitude assessment. Feedback is immediate to teachers and students in the form of visual reports, and while they can be used for many purposes the major use is to monitor teaching and learning.

Paper 4

"Comparing and contrasting models for monitoring learning in three countries"

Ercikan, Kadriye (University of British Columbia, Canada)*

This presentation will review, compare and discuss models for monitoring teaching and learning in three countries that will be presented in the first part of the symposium: Denmark, Hong Kong and New Zealand. The three systems will be reviewed and discussed in terms of their influences on learning and teacher autonomy, the stakes associated with assessments, the types of assessments used, the levels of aggregation of data from these assessments, and how data are used.

Discussant

von Davier, Alina A. (Educational Testing Service, USA)

Monday, 19th July 2010

S01 11:00-12:30 Room211

Cross-cultural examinations of teachers' conceptions of assessment and feedback: Results from survey studies in China, Cyprus, Hong Kong, & New Zealand.

Chair

Brown, Gavin T L (The Hong Kong Institute of Education, Hong Kong SAR, China)

Symposium Abstract

Validity theory in testing and assessment focuses on consequences of testing as an indicator of quality. When classroom testing, scoring, interpretation, and reporting are carried out by school teachers, their beliefs may strongly influence the validity of such practices. There exist a number of persistent tensions around the purposes of assessment (e.g., learning growth vs. student well-being; school/student accountability vs. learning) that influence teacher conceptions and practices of assessment. It is likely that teachers in differing jurisdictions will have differing conceptions of assessment in response to the different priorities put on the various competing purposes of assessment. Hence, cross-cultural comparisons in teacher thinking about the nature and purpose of assessment and feedback can improve our understanding of the validity of assessment. The validity of survey research inventories which were not initially developed for use in cross-cultural research is an issue when making comparisons. Notwithstanding the problems of translation equivalence, such inventories may not fully reflect teacher thinking in societies that have differing priorities for assessment. Hence, cross-cultural research can lead to the identification and measurement of constructs not considered relevant in earlier mono-cultural studies. This symposium reports four survey studies with school teachers in New Zealand, Cyprus, Hong Kong, and China about their conceptions of the nature and purposes of assessment and feedback. These studies provide developments in the validity of school-based assessment by identifying teachers' pre-existing conceptions more fully and shed light as to the relative priority in teachers' espoused beliefs about learning growth, student well-being, and accountability purposes.

Paper 1

"Teachers' Conceptions of Feedback: Results from a national sample of New Zealand teachers"

Brown, Gavin T L (The Hong Kong Institute of Education, Hong Kong SAR, China)*
Harris, Lois R (The University of Auckland, New Zealand)
Harnett, Jennifer (The University of Auckland, New Zealand)

A key step in assessment is responding (i.e., feedback) to interpretations of learner performance. Effective feedback focuses on task, processes, and self-regulation, rather than the self, in other words, it is on a growth pathway rather than well-being. Teachers' beliefs about the nature and purpose of feedback may explain how feedback is implemented. A 71 item Teachers' Conceptions of Feedback inventory was trialled on a nation-wide sample of New Zealand primary and secondary school teachers (N=518). Participants indicated their degree of agreement for each item using a 6-point, positively-packed rating scale. Exploratory factor analysis (MLE, oblimin rotation) retained 48 items in 10 factors. These were tested with CFA in an inter-correlated model, with acceptable fit (χ2=2444.97; df=1035; χ2/df=2.35, p=.12; gamma hat =.90; RMSEA=.051; SRMR=.061). Teachers agreed most highly that feedback focuses on improving student learning (M=4.90), feedback focuses on learning processes (M=4.45), and is interactive between teachers and students (M=4.18). In contrast, teachers rejected the idea that grades were effective feedback (M=2.45), that feedback is done because it is expected by external stakeholders (M=2.88), and that feedback is about accurately grading student work (M=3.06). The notions that feedback is timely, student-led, is about student well-being, and is teacher-only had mean scores between 3.40 and 3.87. These data suggest that that New Zealand teachers' espoused conceptions of feedback lie predominantly on a growth-pathway, rather than the well-being pathway.

Paper 2

"Teachers' Conceptions of Assessment: Cross-cultural testing of models"

Michaelides, Michalis (European University Cyprus, Cyprus)*
Brown, Gavin T L (The Hong Kong Institute of Education, Hong Kong SAR, China)

Surveys of New Zealand and Queensland primary and secondary teachers with the Teachers' Conceptions of Assessment (TCoA) inventory report four hierarchical, inter-correlated factors. That research was conducted only in English in two jurisdictions with very similar policies of low-stakes testing prior to senior secondary school qualifications. Cyprus has a relatively low-stakes assessment policy during the compulsory school years, terminating in high-stakes high-school graduation and university entrance examinations. A functionally equivalent translation into Greek of the 27-item abbreviated TCoA instrument was carried out. A survey of 249 Cypriot teachers (37% <5 years experience) responded using a 6-point, balanced agreement scale, instead of the original positively-packed rating scale. The original Brown model was inadmissible due to very high negative error variance on one 1st-order factor. Removing all 1st-order factors produced an admissible solution with marginal fit, albeit with some very weak regressions. Exploratory factor analysis (MLE, oblimin rotation) of the Cyprus data suggested an alternative five factor solution, which, when modeled as two inversely correlated 2nd-order factors (i.e., assessment is positive and negative; r=-.49), had acceptable fit (χ2=560.51; df=246; χ2/df=2.28, p=.13; gamma hat=.90; RMSEA=.072; SRMR=.073). In a multi-group confirmatory factor analysis, the fit of the Cyprus model to the NZ data showed that the two groups were not statistically invariant. We conclude that small sample size affected the quality of modeling. However, cultural differences in teacher conceptions of assessment are evident. Despite the promise of the TCoA inventory, further item and scale development is required for use outside of Australasia.

Paper 3

"Teachers' Conceptions of Assessment: Developing a model for teachers in China"

Gao, Lingbiao (South China Normal University, China)*
Han, Yuna (South China Normal University, China)
Cai, Ze Jun (South China Normal University, China)

China has a long-standing use of examination-style assessment for competitive student selection to limited educational resources. In this context, teachers have strong awareness that assessment is used to control teacher and student practices and that assessment is fundamentally defined as examinations. In considering the validity of the Teachers' Conceptions of Assessment (TCoA) inventory for use in China, it is apparent that control and examination constructs are not represented. In a small study of Chinese polytechnic instructors, it was found that grading students and evaluating school quality were highly linked with the conception that assessment is for improvement of teaching and learning. A new Chinese-TCoA inventory was developed in the hope of identifying a wider range of constructs that better take into account teacher beliefs in China. A survey of over 1500 teachers in two provinces of China was conducted with a 62 item inventory which expanded the TCoA inventory. A confirmatory factor analysis of the intended 10 factors was inadmissible due to the covariance matrix not being positive definite. An exploratory factor analysis (MLE, oblimin rotation) identified 7 factors. Confirmatory factor analysis of 7 inter-correlated factors, with 39 items, generated acceptable fit ($\chi2$=3442.51; df=539; $\chi2$/df =6.39, p<.01; gamma hat = .90; RMSEA=.059; SRMR=.062). The factor assessment is examination was correlated (.60≥r≥.84) with assessment guides teaching, assessment controls teachers, assessment is for improvement, and assessment controls students. We conclude that the addition of examinations and control are necessary constructs in understanding Chinese teachers' conceptions of assessment.

Paper 4

"Teachers' Conceptions of Assessment: Developing a model for teachers in Hong Kong."

Brown, Gavin T L (The Hong Kong Institute of Education, Hong Kong SAR, China)
Hui, Sammy K. F. (The Hong Kong Institute of Education, Hong Kong SAR, China)*
Yu, Wai Ming (The Hong Kong Institute of Education, Hong Kong SAR, China)

Hong Kong has an assessment for learning policy and a cultural context that emphasizes examinations. In addition to associating student grading with improvement, important improvement-oriented conceptions have been identified among Hong Kong teachers and which were not fully instantiated in the original Teachers' Conceptions of Assessment (TCoA) inventory. An expanded Chinese-TCoA inventory was administered to 601 teachers (85% primary school) in Chinese. The intended 10 factor structure was not supported. Exploratory factor analysis (MLE, oblimin rotation) identified 7 factors. Confirmatory factor analysis of 7 inter-correlated factors, with 33 items, had acceptable fit ($\chi2$=1712.93; df=474; $\chi2$/df =3.62, p=.66) were between assessment is for students' personal development and assessment indicates school quality and assessment is examinations; school quality had similar strength correlations to assessment indicates teacher quality. The conception that assessment is irrelevant had statistically non-significant correlations with assessment for personal development and teachers take account of error; it was also inversely correlated (r=-.31) with assessment improves students' learning. Teachers, on the average, more than moderately agreed that assessment improves students' learning and teachers take account of error; they slightly to moderately agreed with examinations, school quality, and personal development; and disagreed that assessment is irrelevant and is used for teacher quality. We conclude that, in keeping with the highly selective system, Hong Kong primary school teachers predominantly conceived of assessment as examinations for improved learning, rather than as a means of personal development.

Discussant

"Considering the meaning of assessment to teachers across cultures: Insights for the validity of testing"

Kennedy, Kerry J. (The Hong Kong Institute of Education, Hong Kong SAR, China)*

A discussion on the meaning of variations in teachers' conceptions of assessment across the studies in response to differing cultures and policies will be undertaken. How assessment is understood may shed light on how international testing programs are implemented as well as how high-stakes accountability and student qualification assessments are understood and implemented.

S02 14:00-15:30 LT5

Verification models for Unproctored Internet Testing: Strategies, research and considerations for practice

Chair

Weiner, John (PSI Services LLC, USA)

Symposium Abstract

The availability of powerful and affordable technology combined with the need to manage costs has led many organizations to adopt self-service HR systems, which now often include unproctored Internet testing (UIT). While assessment professionals have been debating the merits of UIT, organizations have proceeded to adopt UIT programs. Consequently, the much of the professional debate has moved from whether to how organizations should deploy UIT effectively to realize benefits (e.g., reduced costs and time-to-hire) while managing threats to measurement, particularly cheating. Verification testing has been identified as one strategy for addressing cheating in UIT (Tippins, 2006). There are many ways to cheat on tests whether proctored or unproctored. Key concerns in UIT are: (1) Identification – is it the really the candidate taking the test or a stand-in? (2) Assistance – is the test taker working alone and without unauthorized aids? (3) Exposure – test content can be easily copied and distributed to others, or posted on the web. A variety of approaches to verification testing may be used to address these concerns. However, to date there has been little research or discussion of best practices for models, criteria, or administrative guidance for using the results of verification tests. This session brings together experts from three major assessment firms, who will draw from experience and research to discuss verification testing models, considerations, and research addressing the effectives of alternative approaches to address cheating concerns.

Paper 1

"The utility of verification testing in UIT"

Weiner, John (PSI Services LLC, USA)*

A variety of approaches may be taken to verify candidate scores obtained via UIT. Typically, this entails the administration of a follow-up assessment of some sort under supervised conditions, in order to affirm the candidate's results. While the fundamental concept of verification is simple, there are many variations in approach and use of verification tests, and the practice is not well documented or researched. For example, the verification test may longer or shorter than the UIT, it may measure the same or different constructs, and it may be used as a "cheating detector" to determine whether the candidate's score is sufficiently high to accept the original UIT result, or it may be used in lieu of the UIT results. There are no practice guidelines in this area. This presentation will provide an overview of alternative verification models and research findings on the potential impact of cheating on test validity and decision-making, and will explore the utility of test verification under different scenarios by examining trade-offs and cost-benefits. The presentation will highlight key considerations for those who are contemplating the use of verification testing, and will identify a variety of questions for application and future research. In exploring these issues, research and examples germane to talent assessment will serve as a back-drop.

Paper 2

"From simulations to live data: Does verification work?"

Burke, Eugene (SHL Group PLC, UK)*

This presentation will explore key findings of field research and simulations examining the effectiveness of verification strategies in detecting cheating across various cheating conditions, and whether it is possible to capture and articulate a model for why verification influences the propensity to cheat. Cheating conditions are examined through analyses of live and in vivo data from candidates across Asia, Europe and the US. A two-stage UIT process is described, in which a short verification test operates solely to validate the consistency of candidate scores. The key findings from this paper are that base rates do vary according to factors such as job level which ally with models of the resources available to different types of candidates; that in some cases the rates of non-verified scores are greater than twice the rate expected by chance ad measurement error; that verification processes are very effective in detecting possible cheating by collusion or by proxy, and these processes recover the economic value lost through high false positive rates resulting from cheating; and that the influence of verification in maintaining the integrity of UIT can be explained using a simple model drawing from social psychology and behavioural economics supported by observed candidate behaviours.

Paper 3

"An innovative use of CAT in verification testing"

Fetzer, Michael (Previsor, USA)*

This presentation will examine a method for verification testing that is based on a computer adaptive testing (CAT) model. Results from several case studies will be provided and the value of this approach to verification testing will be presented. This method utilizes the detailed item response information collected during the unproctored session and compares it to that obtained during a short, adaptive assessment. While the test taker is completing the onsite assessment, the algorithm determines whether or not the item response information is consistent and adapts the test accordingly. If consistency is found, the test terminates quickly. If the item response information is inconsistent, the system administers additional items until a consistent pattern emerges and a reliable, valid score can be reported. This innovative process greatly enhances the security of the test items since proctored static tests can be easily compromised and removes the potentially false implication that applicants have cheated. Further, this method also alleviates some of the issues of non-standardized unproctored testing environments by focusing on the consistency of test taker response patterns; it provides applicants the opportunity to do their (personal) best during the hiring process. Research on this new method is currently underway as this process has been recently implemented in several organizations. Full results and case study examples will be provided during the presentation, and the benefits this method brings to the UIT challenge will be highlighted.

S04 14:00-15:30 Room201

Testing, measurement and predictions for personnel decision in Mainland China---An exploration and practice of psychological assessment made by Beisen research

Chair

Zhou, Dan (Beisen Research, China)

Symposium Abstract

Psychological Assessment have been widely used in the area of Personnel Decision in the past decades in China Mainland. There are thousands of companies who are engaging in personnel consulting services, but there are few research on the effectiveness of Psychological Assessment in the area of Personnel decisions. Is there any difference between paper-and-pencil test and computer-based test in personnel assessment? Which psychological measurement format is a better measurement method on theory of types? What is the difference on organizational culture between different types of enterprises? Can we get a pattern classification algorithm in employee selection? What is the relationship among assessment center, 360-Degree, personality, and cognitive ability tests in predicting executive managers? This symposium will give you an point of view from the practice of Beisen Research, and it provide an open forum for researchers from around the world to discuss how to get better effectiveness in applying psychological assessment into personnel decisions.

Paper 1

"Predicting managical performance with 360-degree, personality, cognitive ability and situational simulations tests"

Zhou, Dan (Beisen Research, China)*
Song, Xiaohui
Wang, Maggie

Many methods of 360-degreee, personality questionnaires, cognitive ability Test and situational simulations tests ratings of various performance attributes have proven useful in personnel selection and promotion contexts. As the theoretical or practical value, the Assessment Center method must show incremental predictive accuracy over single method given taking no account of the cost. However, which method can provide predictive value over and above that provided by work experience, level of degree and quality of training? In the present study, we investigated this issue in the context of promotion of executive managers in Chinese departments. Candidates completed a set of 360-degreee, personality questionnaires, cognitive ability tests and a trial of situational simulations tests. The criterion measure was the final grade at the performance of the organization. The prediction of performance of 360-degreee, personality questionnaires, cognitive ability and situational simulations tests were discussed on the following variables: gender, age, number of years of work experience. Results indicated that each method provide unique contribution to the understanding and prediction of performance and were discussed in terms of their implications for the criterion-related and construct validity of Assessment Center. Future research should examine these relationships on a different sample of employees.

Paper 2

"Exploration and practice of ipsative forced-choice measure in talent management"

Li, Bo (Beisen Research, China)*
Liu, Tengfei (Beisen Research, China)
Liu, Qiao (Beisen Research, China)

Psychological measurement have been used extensively in the area of Talent Management, well-known as psychological inventories and scales, in which most of them are normative measures that can be used to inter-individual comparisons. In the past decades, ipsative tests were promoted to be used in recruitment and selection by many publishers who claim that it will give valid insights into the psychological characters of job applications. But many researchers express their strong opposition to such claims. They argued that ipsative tests have its own mathematical prosperities, it cannot be used as normative tests to be compared among individuals. However, this did not mean ipsative tests are of no use in Talent Management. It resolved one of the problems we encountered in the practice of Career Anchor Theory and Test. Career Anchor is a combination of perceived areas of competence, motives, and values that you would not give up (Edgar H. Schein, 1990). Your career decisions will be easier and more valid if you have a clear understanding of your own orientation toward work, your motives and your values. Schein's research on career anchors has shown that there are 8 categories anchors and most people can be described in terms of the 8 anchors presented. The career orientations inventory used Likert item formats(6 points). And we followed this formats when we introduced it into China in 2004. But, in the applications, we found most Chinese individuals choose 5(the statement is often true for you) or 6(the statement is always true for you) which resulted in that they got high scores on all 8 career anchors, we cannot find their real career anchor and give corresponding career development suggestions. The formats of career orientation inventory caused high ratio of people that have not a clear career anchor. In 2006, we developed a new career orientation inventory using ipsative measurement format based on the career anchor theory. It has high reliabilities and validities. The practice in the past three years proved that ipsative forced-choice format is a better measurement method than Likert item formats on theory of types.

Paper 3

"A comparison of paper-and-pencil test and compute-based test in personnel assessment"

Li, Chongliang (Beisen Research, China)*
Zhao, Xiaodi
Zhang, Jiayu

Personnel assessment has been widely used in educational and business settings to help making more accurate evaluation and decision. Traditionally, the main implement form of it is the paper-and-pencil test. Along with the widespread application of computer technology, the computer-based test is being gradually and largely used, and this is going to be the trend in personnel assessment. The two forms possess their own characteristics, however, few studies compared the two types of answering style from a psychometric perspective.

The aim of this research is to systematically examine the difference between paper-and-pencil test and computer-based test in personnel assessment with the following aspects: 1). Personality Test 2). Management In-basket Test 3). Management Scenario Test 4). Ability or Skill Test. Test scores, scale reliability are compared through ANOVA and Correlation Analysis. And also the study examines the variation of these comparison results on demographic and social-economic factors, such as age, gender, educational level, and career. Finally, the practical and theoretical value of this study is discussed in the field of personnel assessment.

Paper 4

"Exploring the possibility of applying pattern classification algorithm in employee selection"

Li, Qi (Beisen Research, China)*
Guo, Jingping

Pattern classification is "the act of taking in raw data and taking an action based on the category of the pattern". A good classifier is particularly helpful, where a classification is needed, but the multidimensional data we are faced exceed the human capacity of extracting patterns. This technique has been used in various areas, such as in computer-aided diagnosis, automatic image detection, and classification of text (for example, spam or non-spam mails). Employee selection is in essence a (dichotic) classification process, because hirers spend a large amount of money and time just to ensure correct classification of applicants into two categories: suitable employees and unsuitable ones. It is common practice now for recruiters to base their preliminary selections on psychological assessment data, but such data are often too huge and complex, and therefore are more bewildering than helpful, if improbably used. The present study aims at exploring the possibility of applying pattern classification algorithm in employee selection to simplify the conversion of psychometric data into usable information, in a real recruitment context. We expect that such an algorithm will make better predictions of whether an applicant will be finally hired, and therefore help hiring companies to make more precise judgment in an earlier stage, so as to cut the cost of recruitment.

Paper 5

"Development and validation of Beisen's Organizational Culture Survey"

Xu, Jiehong (Beisen Research, China)*
Liu, Xueyuan
Tian, Juanjuan
Zhao, Dongyan

There have been many definitions of organizational culture present in the literature. Despite the lack of consensus on the definition, many authors belief that organizational culture can be assessed quantitatively (Turnipseed, 1988). Therefore dozens of instruments have been designed, such as Quinn & Cameron's (1998) Organizational Culture Assessment Instrument (OCAI), Denison's (1995) Organizational Culture Questionnaire (OCQ), and Chatman's Organizational Culture Profile (OCP). As organizational culture is unavoidable related to national culture, these assessments of organizational culture constructed in the Western culture might not be well suitable in the Eastern culture. After a review of the significant studies in this area and most of the research conducted by Chinese scholars, Study 1 in the present paper summarized up 14 dimensions for assessing the organizational culture. In order to bring some structure in, efforts had made to organize these dimensions based on Quinn & Rohrbaugh's (1983) Competing Values Framework (CVF). Exploratory factor analyses and confirmatory factor analyses were conducted to exam the presence of the significant component and the structure of these factors. Except for the internal consistency, inter-rater reliability was also checked in order to find the appropriate level of aggregation, as culture is essentially measured individualistic. Using the survey verified in Study 1, Study 2 further conducted a wide range of investigation into Chinese companies to exam the different culture between different types of enterprises.

S05 15:45-17:15 Room211

Growing talent: The role of assessment in emerging and developed markets

Chair

Harris, William (Association of Test Publishers, USA)

Symposium Abstract

The need for talent drives labor policies of corporations and governments. Corporations seek to acquire and develop world-class workers. Governments seek to create high performance learning environments. Governments grow talent that corporations harvest in pursuit of business objectives. A common thread is assessments. This symposium examines the role of assessments in both developed markets such as Europe and emerging markets. The aim is to contrast current assessment uses and to discuss future applications. China and India are the largest export and domestic markets for goods and services. As domestic and multi-national companies compete for talent, they demand methods that match accurately talent with opportunity. For the European Union (EU) assessments are part of the talent agenda. The EU provides a talent perspective from which to examine the hurdles and opportunities facing China and India. The EU seeks to create a renewable stream of talent to fuel its drive to be the most competitive, knowledge-based economy. As the EU and the growth market countries work to solve the talent challenge, innovative assessments will form part of the solution. Besides examining talent initiatives of governments, a multinational survey of professionals with emphasis on India, provides insight into the corporate view of talent needs and the value of assessment. This session brings together international practitioners and business leaders to discuss talent assessment of today and tomorrow.

Paper 1

"The EU talent pool: Gaining advantage through assessment"

Schuchart, Nadine (Hogrefe Verlag Gmbh & Co., KG, Deutschland)*

A 21st Century goal of the European Union (EU) is to establish itself as the premiere, knowledge-based, digital economy. This goal is aggressive, but attainable. Implicit in this objective is the redesigning of the way knowledge workers (i.e., talent) are assessed. New and innovative assessment approaches are needed to build the requisite world-class talent pool. Becoming the most competitive economy means becoming the leader in all areas of the talent process, including assessments. The next generation of assessments must meet the demands of the next generation workforce. This requires recognizing the changing nature of work, the fluidity of skill sets, the use of creative training methodologies and understanding the critical role of broad competencies (i.e., soft skills) in talent development. These demands require assessments that are seamless, psychometrically robust and that can provide predictive, real-time information about individuals or groups. This future workforce is diverse, virtual, empowered and cognitively nimble. The presentation examines the current ways talent is assessed and discusses ways assessments will be transformed to meet the future EU talent needs.

Paper 2

"Using certifications exams to meet China's talent needs"

Tong, Alex (ATA, China)*

China impressive economic growth has created new demands on its workforce. The Chinese Government has responded with new educational policies and learning environments designed to produce talent to meet the needs of a rapidly evolving and expanding economy. Multi-national corporations in China have partnered with the government to address immediate talent needs. With China's large population and economic successes it is easy to assume that China possesses sufficient talent from its educational system. Chinese labor statistics show a serious shortage of talents, and the number of jobs left unfilled in the skilled job market is substantial. Presently, talent fulfillment needs reach well beyond the capabilities of the existing educational infrastructure of the country. Domestic and multinational companies are turning to training and multi-level certification programs to meet the talent demand. This presentation examines the pivotal role that certification exams play as a supply source of talent to the Chinese economic engine. The developing certification model relies on sophisticated training modules that are layered with assessments so as to monitor the acquisition of knowledge and skills. The presentation includes discussion about the ways assessments will be transformed to meet China's growth beyond the manufacturing sector.

Paper 3

"Talent from a global perspective: A view from 337,000 professionals and managers across 21 countries"

Burke, Eugene (SHL, UK)*

This paper is based on a simple model of performance in which talent reflects the product of potential and experience. This paper is also based on the premise that competencies can be described through "etic" constructs and that the assessments used to measure potential against these competencies are also "etic" in nature where there are empirical data to support that assertion (as will be briefly described in the introduction to this paper). The performance model described has two higher order constructs: execution or the capability to get things done, and engagement or the capability to bring others with you. The data on a large international sample of professionals and managers are used to create an international benchmark against which all countries can be compared. With that benchmark in place, the talent pools in major economies are compared in terms of likely sources of economic advantage and areas where investment in people development is likely to yield yet further economic gains. Four economies in particular are highlighted: China, India, US and Western Europe. The relationship between these profiles and GDP growth for 21 major economies over seven years are also shared, and which clearly show that GDP growth is associated with the talent profiles in the economies covered.

Paper 4

"Thoughts on the direction of talent assessment"

Harris, William (Association of Test Publishers, USA)*

Each of the three presentations will be discussed separately with the aim of highlighting (1) the shift in talent needs for the 21st Century, (2) the advances in assessment and (3) the effects of government policy and corporate demand on the delivery of talent assessment services in a knowledge-fueled, global economy. This review of the presentations will serve as the basis for a higher-order analysis of the future direction of talent assessment services. As market forces reshape all aspects of the worldwide economy some of these forces will touch significantly the ways in which assessments will be developed and delivered. The presentation will use the information from the three presenters to identify emerging trends in talent assessment.

S06 15:45-17:15 LT4

The development and validation of the Multi-axial Clinical Assessment Inventory

Chair

Leung, Freedom Yiu Kin (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

Symposium Abstract

The proposed symposium is about the development and validation of the Multi-axial Clinical Assessment Inventory (MCAI), a clinical instrument derived from the indigenous Chinese Personality Assessment Inventory-2 (CPAI-2). The MCAI is designed according to a coherent theoretical framework of psychopathology. It aims to provide clinicians with comprehensive and clinically useful information on patients' 1) disordered personality features, 2) psychosocial maladjustment, and 3) common psychiatric symptoms. In the first paper, the rationale behind the development of the MCAI will be discussed. Some preliminary analyses about the reliability and validity of the MCAI scales will be presented. In the second paper, item response theory will be used to examine the psychometric properties and clinical utility of the Substance Abuse Scale (SUB) of the MCAI. Preliminary findings support the SUB as a psychometrically sound measure with high clinical utility. In the third paper, the MCAI profiles among psychiatric patients with impulsive personality disorder (IPD) and borderline personality disorder (BPD) will be compared. Preliminary findings support the utility of the MCAI in differentiating IPD patients from BPD patients. In the fourth paper, the MCAI profiles of Chinese adults with problematic hypersexuality (PH), problem gambling (PG), problem drinking (PD) and normal controls will be compared. Preliminary findings support the utility of MCAI in differentiating PH, PG, PD, and normal controls. The implications of these findings for clinical assessment of psychopathology among Chinese psychiatric patients will be discussed.

Paper 1

"The development and validation of the Multi-axial Clinical Assessment Inventory"

Leung, Freedom Yiu Kin (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)*

Cheung, Fanny (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

Fu, Kei (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

You, Jianing (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

Lai, Ching Man (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

Li, Xixi (Department of Educational Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

The present study restructured the Chinese Personality Assessment Inventory-2 (CPAI-2), an indigenous personality test, and developed a new clinical assessment instrument called the Multi-axial Clinical Assessment Inventory (MCAI). The MCAI is designed according to a coherent theoretical framework of psychopathology. It aims to provide clinicians with comprehensive and meaningful information on patients' 1) disordered personality features, 2) psychosocial maladjustment, and 3) common psychiatric symptoms. Two large samples were used in this study. Sample 1 included 1911 non-clinical individuals and sample 2 included 1659 psychiatric patients from Mainland China and Hong Kong. Both samples completed the CPAI-2. Sample 2 also completed the Chinese Personality Disorders Inventory. Results reveal good psychometric properties for all MCAI scales. Convergent and concurrent validity of the MCAI scales have also been demonstrated. Findings also provide preliminary support to the clinical utility of the MCAI psychiatric symptom scales in differentiating patients with different psychiatric diagnoses. The MCAI will provide clinicians in China, Hong Kong and Taiwan a theory-based, culturally relevant, and user-friendly clinical assessment instrument.

Paper 2

"Using item response theory (IRT) in evaluating the psychometric properties and clinical utility of the Substance Use Subscale of the MCAI"

You, Jianing (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)*
Lai, Ching Man (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)
Fu, Kei (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)
Leung, Freedom Yiu Kin (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)
Cheung, Fanny M. (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

This paper used item response theory to examine the psychometric properties and clinical utility of the Substance Use Subscale (SUB) of the Multi-axial Clinical Assessment Inventory (MCAI). A sample of 1,911 non-clinical individuals and a sample of 1,659 psychiatric patients from Mainland China and Hong Kong participated in this study. Exploratory and confirmatory factor analyses supported a unidimensional structure of the SUB. This scale assessed most accurately at moderate to high levels of substance use problems among both samples. Most items discriminated well among individuals with varied levels of substance use problems. The SUB also demonstrated good convergent validity, as suggested by its associations with measures of disordered personality features, such as interpersonal distrust, antisocial behaviors, and attention seeking, as well as with other psychiatric symptoms, such as impulse and anger dyscontrol. Moreover, the SUB could differentiate well between patients with substance use problems and patients with other psychiatric diagnoses. Findings indicate that the SUB of the MCAI is a psychometrically sound measure of great clinical utility.

Paper 3

"MCAI profiles of Chinese psychiatric patients with Impulsive Personality Disorder vs. Borderline Personality Disorder"

Lai, Ching Man (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)*
You, Jianing (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)
Fu, Kei (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)
Leung, Freedom Yiu Kin (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)
Cheung, Fanny (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

This study compared the Multi-axial Clinical Assessment Inventory (MCAI) profiles of Chinese psychiatric patients displaying impulsive personality disorder (IPD; n=72), borderline personality disorder (BPD; n=28), comorbid IPD and BPD (n=25), and neither IPD nor BPD (n=1177). Items from the Chinese Personality Assessment Inventory -2 (CPAI-2) and Chinese Personality Disorder Inventory (CPDI) were selected to assess diagnostic features of IPD and BPD. Results indicated that the four groups differed significantly on disordered personality features, psychosocial maladjustment, and psychiatric symptoms as assessed by the MCAI. Findings indicated the IPD patients showed higher externalizing tendency than BPD patients, the comorbid group showed elevations on almost all of the MCAI subscales than other groups. MCAI was found to be useful in differentiating the IPD, BPD and the comorbid IPD/BPD patients.

Paper 4

"MCAI profiles of Chinese adults with Problem Hypersexuality, Problem Gambling, and Problem Drinking"

Fu, Kei (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)*
You, Jianing (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)
Lai, Ching Man (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)
Leung, Freedom Yiu Kin (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)
Cheung, Fanny (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

The present study examined the MCAI profiles of Chinese adults with problem hypersexuality (PH), problem gambling (PG), and problem drinking (PD). Chinese adults with problem hypersexuality (N = 42), problem gambling (N = 36), problem drinking (N = 70) and normal controls (N = 95) were compared in this study. Findings revealed distinctive MCAI profiles among Chinese adults with PH, PG and PD. These deviant groups scored significantly higher in externalizing personality features such as antisocial behavior and attention seeking, and psychiatric symptoms such as substance use, impulse and anger dyscontrol. The PH group also showed significantly more problems in psychosocial maladjustment than other groups. The PH and PD groups reported more internalizing symptoms than the other two groups, especially in the anxiety and somatic complaints. This study provided preliminary support to the clinical utility of the MCAI in differentiating individuals with PH, PG and PD. MCAI revealed clinically meaningful differences in disordered personality features, psychosocial maladjustment and psychiatric symptoms among these deviant groups. This information may be useful in helping clinicians to make more accurate assessment of the underlying psychopathology of these deviants individuals.

Discussant

Leung, Freedom Yiu Kin (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)*

Implications of these findings for psychological assessment among Chinese psychiatric samples will be discussed.

Tuesday, 20th July 2010

S03 08:30-10:00 LT3

Reforming assessment and examination for the 21st century in a city of East meets West

Chair

Cheung, Kwok Wah (Curriculum Development Institute, Education Bureau, HKSAR Government, Hong Kong SAR, China)

Symposium Abstract

China has started examination for entering civil service back to the Han Dynasty, and this practice has passed from one generation to another. To respond to the substantial challenges of the 21st Century and to enable our students to develop learning to learn capabilities, Hong Kong has started a holistic review on the school curriculum as well as assessment practice. With the curriculum reform implemented in 2001 and the New Senior Academic Structure in place since 2009, assessment must be changed to ensure alignment with the structural and curriculum changes. Learning and assessment complement each other. As curriculum reform focuses on the development of students' generic skills, assessment reforms have to be made to ensure that student performance in these areas can be properly assessed and accredited. The assessment system should also be designed to cater for the needs of a wide ability range of students. The theme of the HK symposium is therefore on "Reforming Assessment and Examination for the 21st Century in a city of East meets West", with 4 sub-themes where we could bring out how we balance the perspectives of designing modern assessment and explain the challenges we have now. These sub-themes are:

1. From Ancient Country to Modern Hong Kong - the historical-cultural contexts and assessment reform framework
2. Assessment for learning: supporting curriculum reform
3. Territory-wide System Assessment as effective feedback for teaching and learning
4. Sharing of school-based assessment policy - the case of a primary school

Paper 1

"From Ancient country to modern Hong Kong -- the historical-cultural contexts and assessment reform framework"

Cheung, Kwok Wah (Curriculum Development Institute, Education Bureau, HKSAR Government, Hong Kong SAR, China)*

China, beginning in the Han Dynasty, started the imperial examinations in which local officials would select candidates to participate in an examination of the Confucian classics. Under this system, any successful candidate could move from a low social status to political prominence as a high-ranking official. In such historical-cultural contexts where the Confucian values are passed until the old examination was abolished in 1906, we encountered both opportunities and challenges in our assessment reform. One of our challenges in Hong Kong is that schools have inherited an idea of preparing students for public examinations, and such an examination-oriented culture has to be changed. In 2001, proposals to reform the school curriculum with students' interests as our top priority were launched. Besides, a culture of "Assessment for learning" has to be nurtured in schools. Assessment is an integral part of the curriculum, pedagogy and assessment cycle, and it helps to provide quality feedback to different stakeholders and the education system. Another challenge lies in the assessment modes. The reform proposals for basic education encourage schools to collect evidence of student learning by using different modes of assessment suited to the purposes and processes of learning. With the New Academic Structure implemented in 2009, a wider range of approaches to assessment and reporting including the use of school-based assessment, standard-referenced reporting and a student learning profile are introduced. The aim of Hong Kong education is to enable students to build a broader knowledge base for whole-person development and life-long learning.

Paper 2

"Assessment for learning: Supporting curriculum reform"

Cheung, Kwong Yuen, Thomas (Hong Kong Examinations & Assessment Authority, Hong Kong SAR, China)*

In 2001, the reform in our school curriculum stated that students should be provided with essential life-long learning experiences for whole-person development. Students should learn how to learn through developing generic skills. It is also recommended that priority should be given to critical thinking, creativity and communication skills as they are crucial for helping students to learn how to learn. The Hong Kong Examinations and Assessment Authority appreciates the fact that assessment practices have to respond to the requirements in the curriculum reform. Assessment reforms have to be made to ensure that student performance in areas like generic skills can be properly assessed and accredited, through the assessment of understanding in various subject disciplines. Assessment entails setting questions which test understanding by posing to students a novel topic, a theme or a context and assessing what meaning students can derive from the situation and how they utilize information and apply knowledge. Also, an appropriate mode of assessment should be used. Students should be familiar with the requirements of the assessments, and the assessment system should be designed to cater for the needs of a wide ability range of students. A quality assessment reform should therefore focus on the following:

(a) providing authentic assessments
(b) setting questions which emphasise applications
(c) choosing comprehensive assessment schemes
(d) making requirements explicit
(e) making allowance for the varying abilities of learners

This presentation aims to explain the international trends along each of the areas and how they affect the development of local assessment practices.

Paper 3

"Territory-wide System Assessment (TSA) as effective feedback for teaching and learning"

Lam, Ling Chi, Tenny (Hong Kong Examinations & Assessment Authority, Hong Kong SAR, China)*

To improve the Hong Kong education system, the Education Commission recommended a System Assessment (renamed Territory-wide System Assessment later) in Chinese, English and Mathematics (CEM) at P3, P6 and S3 in 2000. The purpose is to provide the Government and school management with information on the school standards in key learning areas so that the Government would be able to provide support to schools in need of assistance. Also, the results from TSA at the territory-wide level would help to monitor the effectiveness of education policies. TSA began at P3, P6 and S3 in 2004, 2005 and 2006 respectively. From 2006 onwards, all students at P3, P6 and S3 take part in TSA. As a low-stake assessment, TSA is not the test instrument for Secondary School Placement Allocation and it will not provide schools data on individual student. Each of the items set in CEM subjects is based on a series of Basic Competency descriptors which have a close link with teaching and learning. With the results reported to individual school, the schools would be able to use them in curriculum planning and learning and teaching so as to enhance their effectiveness. TSA has been implemented for years and its impact on learning and teaching in schools is worth studying. The symposium will focus on the rationale of TSA, the standards maintained across different years and levels, the reporting mode and channels as well as the surveys conducted which help to review the effectiveness of TSA.

Paper 4

"Sharing of school-based assessment policy - the case of a primary school: Baptist (STW) Lui Ming Choi Primary School"

Tang, Mei Shin (Baptist (STW) Lui Ming Choi Primary School, Hong Kong SAR, China)*

Baptist (STW) Lui Ming Choi Primary School has profound experience in school-based assessment. As early as 1992, the School recognized the significance of including 'assessment for learning' as an integral part of the 'teaching-learning–assessment' cycle and started school-based assessment in its curriculum integration development. To enhance students' whole-person development and learning capabilities, the School formulated a school-based assessment policy that emphasised assessment for learning shortly after the introduction of the curriculum reform in 2001. A better balance between 'assessment for learning' and 'assessment of learning' among various Key Learning Areas (KLAs)/subjects has been achieved. For instance, the School has developed School-based Assessment Indicators in the Chinese Language and English Language Education KLAs with a view to systematically gauging and reflecting on how to improve student learning. For other KLAs/subjects like Visual Arts, Music and Physical Education, diversified modes of assessment are adopted to effectively assess students' performance in knowledge, skills and attitudes. The School has also strategically reduced the number of summative assessments, reviewed and enhanced the use of formative assessments. To enhance their teachers' assessment literacy and professional capacity, it has tapped professional support strategically from the Education Bureau and tertiary institutions. The School will continue to explore a wider use of 'authentic assessment' and develop an 'i-portfolio' to sustain the positive impacts and good practices. This presentation aims at sharing the experience of developing a school-based assessment culture and incorporating 'assessment for learning' in the whole-school curriculum planning to foster student learning and school improvement.

Discussant

Chan, Ka Ki, Catherine (CQA Branch, Education Bureau, HKSAR Government, Hong Kong SAR, China)*

S07 08:30-10:00 LT5

A combined emic-etic approach to personality assessment: CPAI-A's standardization and application among Chinese adolescents

Chair

Fan, Weiqiao (Department of Psychology, Shanghai Normal University, China)

Zhou, Mingjie (Institute of psychology, Chinese Academy of Sciences, China)

Symposium Abstract

The Cross-cultural (Chinese) Personality Assessment Inventory for Adolescents (CPAI-A) was developed using a combined emic-etic approach which includes universal and culturally relevant personality constructs. The CPAI-A research program aims to establish not only the reliability and validity of an indigenously derived assessment measure, but also to promote understanding of personality beyond that of a universal personality structure. This symposium presents the latest research on the CPAI-A in Chinese settings. The papers report the findings of the standardization studies, and analyze the contributions of universal personality scales and indigenously derived personality scales in understanding mental health, psychological well-being, and behavior problems among Chinese adolescents.

Paper 1

"A preliminary comparison between the standardization studies of the CPAI-A in Hong Kong and Mainland China"

Fan, Weiqiao (Department of Psychology, Shanghai Normal University, China)*

Cheung, Fanny M. (Department of Psychology, Chinese University of Hong Kong, Hong Kong SAR, China)

The CPAI-A was standardized among Hong Kong adolescents in 2005 and among mainland China adolescents in 2008. We compared the factor structures of the personality inventories and clinical inventories of the CPAI-A between the two standardized samples, respectively, and found the results in Hong Kong sample were largely consistent with those of mainland sample. The reliabilities of the subscales of the CPAI-A in both samples were basically statistically accepted: for personality subscales, the average Cronbach α coefficients were .72 for the Hong Kong sample, and .70 for the mainland sample; for clinical subscales, the average was .77 for the Hong Kong sample, and .76 for the mainland sample. We also compared the T scores of all subscales of the CPAI-A between the two samples. Some interesting results were found. Concerning the general personality scales, male adolescents from mainland China scored higher on scales of Sensation-seeking (SEN), but lower on the scales of Discipline (DIS) than their counterparts from Hong Kong; in contrast, female adolescents from mainland China scored lower on the scale of SEN, but higher on the scale of DIS than their counterparts from Hong Kong. Concerning the clinical subscales, male adolescents from mainland China scored higher on the scales of Hypomania (HYP), Need for Attention (NEE), Pathological Dependence (PAT), and Eating Disorder (EAT) than male adolescents from Hong Kong; however, females from mainland China scored lower on the scales of PAT and EAT.

Paper 2

"How disaster experience influenced adolescents' mental health: Findings from CPAI-A"

Zhou, Mingjie (Institute of psychology, Chinese Academy of Sciences, China)*

Fan, Weiqiao (Department of Psychology, Shanghai Normal University, China)

Cheung, Fanny M. (Department of Psychology, Chinese University of Hong Kong, Hong Kong SAR, China)

As part of the standardization study of the Cross-cultural [Chinese] Personality Assessment Inventory for Adolescents (CPAI-A) in mainland China, we collected adolescent samples from the 2008 Sichuan earthquake area at two specific time points:Half a year after disaster and a year after disaster. A total of 1117 adolescents from grade 7 to grade 12 completed the CPAI-A. In addition, the participants' exposure to the disaster was assessed with 8 items. The results showed that (1) compared with those from non-disaster area, those who have disaster experience showed more psychological problems; (2) half a year after the earthquake, exposure to the disaster had a small effect on adolescents' mental health; (3) as time passes, their mental health improved greatly. The results give us great insight in understanding how disaster experience influences adolescents' mental health. The results also showed that the CPAI-A has good ecological validity.

Paper 3

"Smoking among Hong Kong Chinese adolescents: The roles of personality and gender"

Wan, Sarah Lai Yin (The Open University of Hong Kong, Hong Kong SAR, China)*

Personality and gender are important individual factors in predicting adolescent smoking. However, previous behavioral models are often found to be inadequate in explaining the smoking intention among adolescents as individual factors have not been taken in consideration. This study tested the applicability of a Theory of Planned Behavior Model that incorporates individual cognitive-motivational factors, personality factors (Sensation Seeking, Emotional Instability and Interpersonal Relatedness), and family and peer smoking in understanding smoking intention among Hong Kong Chinese adolescents. In a sample of 1332 adolescents, individual attitudes, subjective norms, and perceived behavioral control over smoking, together with personality factors, and family and peer smoking were found to be associated with smoking intention in SEM analyses. Multi-sample invariance analyses also revealed differences in (1) the effects of parental norms, Interpersonal Relatedness, and family and peer smoking on smoking intention across gender, and (2) the effects of smoking attitudes, parental norm, perceived behavioral control, Sensation Seeking, and Emotional Instability on smoking intention across ever-smokers and never-smokers. Findings of the study provided theoretical implications that, together with gender, both universal and culture-specific personality factors had significant influences on smoking intention in explaining adolescent smoking. The practical significance to the design of prevention programs was also discussed.

Paper 4

"The influence of interpersonal relationship on psychological problems among Chinese adolescents: The preliminary study"

Cao, Hui (Department of Psychology, Chinese University of Hong Kong, Hong Kong SAR, China)*
Cheung, Fanny M. (Department of Psychology, Chinese University of Hong Kong, Hong Kong SAR, China)
Fan, Weiqiao (Department of Psychology, Shanghai Normal University, China)

The study examined the relationships among personality and psychological problems, especially how the indigenous measures of personality add the contribution. A total of 1281 secondary students were collected from 10 middle schools in Mainland China with Cross-cultural (Chinese) Personality Assessment Inventory – Adolescents Version (CPAI-A). Two personality factor – Emotional Stability and Interpersonal Relatedness and two clinical indices (Emotional Problem & Interpersonal Problem) which were grouped by five clinical subscales were used in present study. Results of General Linear Modeling showed: For both clinical indices, (1) the main effect of both Emotional Stability and Interpersonal Relatedness were significant; (2) the interaction between Emotional Stability and Interpersonal Relatedness was significant. Results suggested that for Chinese adolescents, beyond emotional stability, the indigenous personality factor - interpersonal relatedness should be paid more attention to. How culturally-relevant personality traits affect Chinese adolescent's mental health was discussed.

S08 08:30-10:00 LT6

Application of computerized adaptive testing

Chair

Chen, Po-Hsi (Department of Educational Psychology and Counseling, National Taiwan Normal University, Taiwan)

Symposium Abstract

Computerized adaptive testing (CAT) had been put into practice in the recent years. However, some problems of using the CAT still need to be solved. For examples, how to enhance the accuracy of items calibration, how to control the item exposure rate and test overlap rate, what are the influences of the constraints of content balancing on ability estimation, and how to detect the cheating behavior. In this symposium, we will discuss these application issues of the CAT. In the first research, a computerized adaptive pretest procedure had been suggested to improve the accuracy of items parameters calibration, especially when only few subjects can be obtain in pretest. In the second research, two methods of controlling the pairwise test overlap rate had been used in CAT to compare their performance. In the third research, the constraints of content balancing had been used with two item selection rules in computerized classification testing and the accuracy of classification were compared. In the fourth research, response time had been used to detect the cheating behavior of the pre-knowledge subjects. The applications and suggestions of these four studies on CAT will be addressed in this symposium.

Paper 1

"Specifying optimum items to the examinees for item parameter calibration in pretest: The computerized adaptive pretest"

Wu, Chia-Ju (Department of Educational Psychology and Counseling, National Taiwan Normal University, Taiwan)*
Chen, Po-Hsi (Department of Educational Psychology and Counseling, National Taiwan Normal University, Taiwan)

The goal of the research is to investigate the influences of the examinee's latent trait on the accuracy of pretest items calibration. The precision of latent trait estimation in computerized adaptive test (CAT) depends on the accuracy of item parameters. However, it may be difficult to get large sample size for pretest items calibration due to the security reason or the lack of appropriate examinees. The authors proposed a computerized adaptive pretest (CAPT) method for specifying the "optimum items" for each examinee in the pretest stage in order to improve the accuracy of item calibration. Two simulation studies were carried out in the research. In study one, three test forms with different items difficulties were calibrated using three populations of examinees with different ability levels. In study two, pretest for items calibration were carried out using the nonequivalent group with anchor test (NEAT) design or CAPT method. The dependent variables were the bias and the root mean square error (RMSE) of items parameters. Results of study one indicated that when the abilities levels of examinees match the difficulties of test forms, the bias and the RMSE were lower than when the abilities levels of examinees didn't match the difficulties of test forms. The smaller the sample size the larger the difference between these two conditions. Results of study two demonstrated that when using CAPT method, the bias and RMSE of items parameters are smaller than using the NEAT design. Suggestions and applications of CAPT were discussed in this research.

Paper 2

"*A comparison of pairwise test overlap control methods in computerized adaptive testing*"

Chao, Hsiu-Yi (Department of Psychology, National Chung-Cheng University, Taiwan)*
Chen, Shu-Ying (Department of Psychology, National Chung-Cheng University, Taiwan)

When item sharing between pairs of examinees is considered, several procedures could be used to control pairwise test overlap. In addition to the SHGT procedure (with) proposed in the first study, Chen, Lei and Liao (2008) developed an online version of the Sympson and Hetter procedure with test overlap control (SHTO), while a simplified version of van der Linden and Veldkamp procedure with test overlap control (SLVT) was proposed by Lin (2007). The purpose of this study is to investigate the similarity among these procedures on pairwise test overlap control. Preliminary results indicated that the SHGT procedure performed identically to the SHTO procedure when a pre-specified maximum test overlap rate was not stringent but outperformed the SHTO procedure when the pre-specified maximum test overlap rate was stringent. Detailed comparison among these procedures will be completed by April, 2010. Based on the results observed from this study, the performance of these procedures on test overlap control can be evaluated, and guidelines for selecting appropriate test overlap control methods when pairwise item sharing is of practical concern can be provided.

Paper 3

"Effects of practical constraints in item selection rules on computerized classification testing"

Hsu, Chia-Ling (Assessment Research Centre, The Hong Kong Institute of Education, Hong Kong SAR, China)*
Wang, Wen-Chung (Assessment Research Centre, The Hong Kong Institute of Education, Hong Kong SAR, China)
Chen, Shu-Ying (Department of Psychology, National Chung-Cheng University, Taiwan)

There are two common item selection rules in CCT: the cut-score based sequential probability ratio test (denoted as CB/SPRT) and the trait estimate based sequential Bayes procedure (denoted as EB/SBP). It has been found that the former method required fewer items than the latter when the same level of classification accuracy was achieved. Whether this advantage holds when practical constraints of content balancing and item exposure control are implemented remain unknown. To investigate this issue, we implemented procedures for content balancing and item exposure control onto the CS/SPRT and the ES/SBP, and evaluate their performances through a series of simulations. The results indicated that (a) the CB/SPRT was more efficient than the EB/SBP when no practical constraints were implemented; (b) these two methods performed similarly when practical constraints were implemented; (c) the lager the item pool was, the smaller the mean test length and the mean test overlap rate would be, and the larger the number of items used and the higher the classification accuracy would be; and (d) under the Rasch model, almost all items in the pool were used, expect for the EB/SBP. In conclusion, if practical constraints are implemented, then the choice between the CB/SPRT and the EB/SBP makes little difference.

Paper 4

"Investigating response times for examinees with item pre-knowledge in computer-based testing"

Chen, Jyun-Hong (Department of Psychology, National Chung-Cheng University, Taiwan)*
Chen, Shu-Ying (Department of Psychology, National Chung-Cheng University, Taiwan)

In computer-based testing, item response times (RT) are easily recorded. RT can provide additional information for the investigation of test items and examinees. van der Linden & van Krimpen-Stoop (2003) used RT to detect cheating behavior of examinees. Cheating behavior means that examinees answer items correctly based on their item pre-knowledge, rather than on their ability. Their study showed that RT was sensitive to cheating behavior. However, they assumed that RT of cheating behavior was followed lognormal distribution. If the assumption was not held practically, the results may not be valid. The purpose of this study is to investigate the properties of RT of cheating behavior. In order to obtain the RT data in practice, we held the MATH-CBT. We manipulated that some items of MATH-CBT were previewed by examinees before test administrating. Therefore, some of the examinees' RT were answered by normal behavior, and some by cheating behavior. After analyzing these RT data, we found that the differences in RT between normal behavior and cheating behavior were significant. We also found that the lognormal distribution was the best fit distribution of cheating behavior's RT. And further analysis of mixture model also proved that mixture model was a powerful tool in detecting cheating behavior within items.

S09 10:30-12:00 LT2

Partnering in meeting assessment needs in China: A case of East and West working together

Chair

Burke, Eugene (SHL Group Ltd., UK)
Tong, Alex (ATA Inc., China)

Symposium Abstract

This symposium describes a case study of two companies, ATA and SHL, operating in different verticals and geographies working together to meet a pressing assessment need in China. The project was to provide a short competency-based assessment for identifying the fit of candidates to jobs organised into several job families. The requirements were or a short, efficient but psychometrically sound tool organised around a clear taxonomy of behavioural competencies to compliment a taxonomy of technical competencies already identified for over 600 jobs ranging from managerial and supervisory levels to skilled and semi-skilled operational roles, and covering roles within information technology through to marketing and sales as well as administration. The symposium describes the assessment need, how the assessment technology originally developed in Europe was adapted for the Chinese market, how the solution was deployed and experience since original deployment. Key learnings from the project are also described including how the project has served to show the importance of softer behavioural competencies to companies in the growing Chinese economy.

Paper 1

"An overview of the assessment need"

Tong, Alex (ATA Inc., China)*

The Chinese job market is facing a huge challenge with around 12 million job openings every year waiting to be filled. At the same time there are a total of around 24 million people actively looking for jobs including 7 million new university graduates from universities without any substantial job experience or job skills. All of these factors make the Chinese job market extremely competitive from a supply side perspective. From the demand side perspective of employers, organizations are becoming more cautious in their hiring decisions due to the increasing costs of errors in hiring given recent Chinese labor laws. ATA recently launched an online talent application to help HR managers and hiring managers tackle this problem. Using competency profiles for over 600 jobs organized by job families, the ATA application enables recruitment managers to handle a large volume of job applications efficiently and to be able to objectively select from hundreds of applicants. The application combines assessments of both critical technical skills and knowledge, and softer behavioral competencies to provide a fit score for each job applied for. Working closely with SHL, an efficient questionnaire has been developed in Chinese to sift candidates against fit to the behavioral competencies for a given job.

Paper 2

"Addressing the need for a competency-based assessment tool"

Burke, Eugene (SHL Group Ltd., UK)*

The requirements established with ATA were for an assessment of fit to valid behavioural competencies for over 600 jobs in Chinese. This assessment had to be efficient to sit alongside assessments of technical skills and knowledge developed by ATA, and to provide users of the assessment application to identify quickly the fit of an applicant to a specific job or a specific job family for the purposes of short listing job applicants. Another key requirement was that the assessment should manage both faking good as well as security issues as the assessments would be delivered online. To meet these requirements, the assessment, a short ipsative questionnaire previously validated in Europe and in several languages, was localised into Chinese and competency profiles for each of the 600 plus jobs was developed using SHL's Universal Competency Framework (UCF) mapped to the US ONET job analysis database. This paper describes the procedures used to develop draft UCF profiles for jobs using subject matter experts, and to refine and validate the final profiles delivered to ATA to provide a coherent taxonomy of job profiles. The psychometric properties of the questionnaire including how it manages faking and security issues are also described.

Paper 3

"What the data tells us about China's talent pools"

Liu, Ying (ATA Inc., China)*

Burke, Eugene (SHL Group Ltd., UK)

This paper describes two sets of analyses. The first is an item and scale comparison of the behavioural competency questionnaire undertaken to ensure that the questionnaire was free of any substantial bias in its Chinese language form. The second series of analyses looks at the overall competency profile for job candidates tested to date and the levels of fit identified for the over 600 jobs currently catered for by the ATA application. These analyses show where the Chinese labour market currently offers potential strengths in supplying talent to meet the needs of organisations, and where there may be areas requiring further development either through educational programmes prior to entering the labour market, or through the investment of employers in the development of those they hire on the basis of best fit to a job.

Paper 4

"Key learning from the project"

Tong, Alex (ATA Inc., China)*

Burke, Eugene (SHL Group Ltd., UK)

This symposium will have described actual experiences of two companies working together to meet a market need in China, but working from different spaces within the testing and assessment industry, and operating in very different geographies and cultures. So, what did these two companies learn about how companies from the East and West could do to work more effectively? This paper closes the symposium by comparing experiences and identifying the practical steps that our mutual experiences suggest companies from the East and West could consider to ensure success. This part of the symposium will also provide time for the audience to ask questions about both operational and technical issues.

S10 10:30-12:00 LT5

Overcoming technical problems with computer-based testing

Chair

von Davier, Alina A. (Educational Testing Service, USA)

Symposium Abstract

There can be no doubt that the next generation of testing will see more use of computers and the Internet for assessing aptitude, achievement, and personality. Tests can be more flexibly scheduled, score reporting can be immediate, and many new important skills can now be measured. At the same time, it would be wrong to suggest that it is simply a matter of moving paper and pencil tests to the computer. Often items calibrated in one context do not function in the same way in another. Also, field testing can be more demanding in a computer-based testing design. New security problems arise too. Items are being exposed to candidates constantly, and the question is, how can the impact of item exposure on test validity be minimized? Other topics under study include the choice of test design, item bank size, and item selection strategies. In this symposium, one goal is to introduce attendees to a host of problems and possible solutions that arise when attempting to use computer-based testing, in practice. Three of the papers will address important problems such as prescreening items for computer-adapt testing, test designs for assigning candidates to performance classifications, and IRT model selection, test lengths and item bank quality for diagnostic testing. A second goal is to focus on some current applications. In the first, we will see the comparative impact of paper and pencil versus computer-administered versions. In the second, the issue of test security with computer-adaptive testing in personnel selection will be considered.

Paper 1

"Developing statistical screening criteria for pretest items on a computerized adaptive test"

Gorham, Jerry (Pearson, Inc., USA)*
Woo, Ada (National Council of State Boards of Nursing (NCSBN), USA)

Developing appropriate pretest screening criteria for a CAT program is an important aspect of maintaining effective and well-supplied CAT pools. One unique characteristic of most CAT pretest designs is that new items are pretested randomly on a wide range of examinee ability but then used on a live CAT exam in a targeted manner for a relatively narrow range of examinee ability. This characteristic becomes even more prominent for variable-length CATs in high information regions on the ability scale such as those near the passing standard. This study will evaluate the effectiveness of certain classical and IRT-based pretest statistics for selecting items that will perform well on a live CAT exam. Outcome criteria for acceptable item performance are based on item-model fit, item discrimination indexes, and stability of the estimated difficulty parameters. The study will examine criteria for both traditional multiple-choice item formats and newer item formats such as multiple response items.

Paper 2

"Computerized classification testing in two or three categories by sequential statistical testing"

Eggen, Theo (CITO, Netherlands)*

When classification into a limited number of categories is the main purpose of testing, algorithms based on the application of sequential statistical testing have shown to be better performing alternatives above traditional estimation based computerized adaptive tests (e.g. Reckase, 1983; and Eggen & Straetmans, 2000). In these studies, the sequential probability ratio test (SPRT; Wald, 1947) is applied in order to decide whether more observations on items are needed and which classification decision is to be made. In case of one cutting point where a fixed optimal test can be defined, recently Finkelman (2008) proposed additional stopping rules to the SPRT. This stochastically curtailed sequential probability ratio test, or SCSPRT, adds some rules by which testing is stopped in case a change in the decision between categories is possible but very unlikely. In the paper the problems encountered and the solutions in generalizing the application of the SCSPRT to problems with more than two categories will be presented. In this case the (optimal) composition of the test cannot be fixed in advance, which is a requirement of the SCSPRT. The performance of the proposed procedures is illustrated by results of simulation studies with item banks calibrated with the one and the two parameter logistic test model (van der Linden and Hambleton, 1997).

Paper 3

"Impact of IRT model selection and test length on computer-adaptive testing"

Yoo, Hanwook (University of Massachusetts, Center for Educational Measurement, USA)*
Hambleton, Ronald K. (University of Massachusetts, USA)

Recently, many education agencies in the US have become interested in computer-adaptive testing (CAT). Moreover, because of the No Child Left Behind act, diagnostic testing is becoming increasingly important to schools while at the same time, schools want to reduce the amount of extra testing time for students. In principle, because CAT can shorten testing time with little or no loss in measurement precision or decision-making capabilities, it would seem to be a promising tool for diagnostic testing. For student diagnostic testing—CAT can be used at the classroom level, and feedback to students can be fast, and without much involvement from teachers. As a prior study (Yoo & Hambleton, 2009) proved, when the test with CAT design is relatively short, student proficiency estimation will likely be quite different with the various item response theory (IRT) models, certainly problematic for diagnostic testing purposes. As shorter tests with a CAT design are becoming more common, the question of model choice seems to be worthy of research. Furthermore, manipulating the quality of the items can be considered by test developers as an important variable for CAT design–not affecting the content of the items but the statistical characteristics of the test items in test design. The purpose of this study was to investigate the capability of dichotomous IRT models to estimate the examinee proficiency in short length CATs. Variables of interest included model selection, test length, and item bank quality. The study was carried out using computer simulation procedures.

Paper 4

"Validity of CAT in personnel selection"

Fetzer, Michael S. (Advanced Assessment Technologies, USA) *

The advent of online testing has brought with it tremendous advantages, but also some very real challenges. Of paramount importance is the issue of test security, especially as assessments are delivered in an unsupervised setting across the globe. In addition, enhancements to the validity and accuracy of online assessments are of great interest to those who utilize these tools for selection and placement. Recently, computer adaptive testing (CAT) has emerged as the best solution to these challenges and is quickly becoming standard practice, especially for unsupervised assessment programs. This presentation will review the development of CAT-based cognitive ability and personality assessments as well as the positive results of recent research indicating their utility in a personnel selection context. Researchers and practitioners alike will benefit from this discussion of a viable new method of pre-employment testing.

Paper 5

"Internet administration of pre-academic preparatory program admissions tests (MEIMAD)"

Gafni, Naomi (National Institute for Testing and Evaluation, Israel)*
Blum, Nadav (National Institute for Testing and Evaluation, Israel)
Baumel, Michal (National Institute for Testing and Evaluation, Israel)

This paper focuses on the Internet administration of the MEIMAD test taken by candidates for pre-academic preparatory programs in Israel. From 1991 to 2008, the MEIMAD test was administered in paper-and-pencil format. Since May 2008, it has been administered solely via the Internet. The test covers three subjects: Hebrew, mathematics, and English. All items are multiple-choice. All aspects of the test's administration are Internet-based, with no reliance on a local server. Thus far, 11,442 candidates for 35 programs have taken the test in this manner. Candidates receive a test form that is randomly selected from a bank consisting of several forms. Results obtained in the two modalities are quite similar in terms of both reliability and average difficulty. The conclusion reached is that Internet administration of the test is efficient, both as a means of administration and for score reporting, facilitating the utilization of existing infrastructure at NITE and in the colleges. Complications encountered during test administration are negligible.

S11 10:30-12:00 Room211

(i)Development and psychometric assessment of a psychopathology measure for use in Mainland China: Adjustment Scales for Children and Adolescents (Chinese Mandarin Translation)

Chair

Ding, Yi (University of Toledo, USA)

Symposium Abstract

This symposium presents a collection of papers detailing the development and psychometric evaluation of a Chinese Mandarin translation of the Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993) for use in mainland China. The ASCA is a teacher-report psychopathology measure (broad-band) standardized with a representative U.S. sample of youths ages 5 through 17, with strong psychometric support. A review of existing Chinese versions of instruments for general or specific child psychopathology revealed that instruments measuring overall adjustment in school settings have yet to be developed and thus highlights the need. Paper 1 details the methods and procedures of translation, back-translation, and consideration of cultural relevancy of ASCA items and dimensions in addition to data collection procedures in field testing that produced a large sample of Chinese children (N = 554) in first through sixth grade (ages 6-13). Paper 2 concerns differential item functioning comparisons between the Chinese sample and comparably aged and grade placed children from the ASCA U.S. standardization sample, as well as across sex and development. Paper 3 examines the factorial invariance of the ASCA to determine if the items were measuring similar dimensions for the Chinese sample compared to the ASCA standardization sample. Paper 4 details examination of scale (core syndromes, supplementary syndromes, overall adjustment scale) differences between the Chinese children and children in the ASCA standardization sample as well as comparisons of item prevalence estimates for ASCA positive behaviors, the 20 least common behavior problems, and 20 most common behavior problems.

Paper 1

"Translating the Adjustment Scales for Children and Adolescents into Chinese Mandarin and field testing data collection in Mainland China"

Ding, Yi (University of Toledo, USA)*
Canivez, Gary (Eastern Illinois University, USA)
Kuo, Yi-Lung (University of Iowa, USA)
Guo, Jian-peng (The University of Hong Kong, China)

China's population of children from birth to 15 years was greater than the overall population of the United States, however, studies of well-validated standardized measures of child psychopathology for the Chinese population are sporadic and limited. The development of a Chinese version of the Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993) to measure the psychological difficulties is described. Test translation procedures included forward translation, backward translation, and comparison by independent researchers and then revised where necessary. Issues and challenges associated with cross-cultural psychological assessment were addressed with respect to the potential bias at the item level, misinterpretation of particular behavioral manifestation, which is culturally specific, and issues regarding linguistic, concept, and metric equivalence. Problematic items in the Chinese ASCA are discussed in detail. Once translation was completed, it was field tested with an independent sample of 554 Chinese students, ranging from first grade to sixth grade in an inner city elementary school. Description of the sampling and data collection for the Chinese ASCA is also reported. Disregarding the long-lasting argument against the importation of measures from other cultures, the authors are in support of the perspective that psychopathological phenomena are basically universal, although considerably influenced by the social-cultural context in which they occur. Based on such perspective, there is neither a disorder entirely free from cultural shaping, nor one that can be entirely ascribed to social or cultural characteristics. Independent rating of relevancy of item content for the Chinese translation of ASCA is noted.

Paper 2

"Adjustment Scales for Children and Adolescents differential item functioning: Chinese and U.S. standardization sample comparisons"

Kuo, Yi-Lung (University of Iowa, USA)*
Ding, Yi (University of Toledo, USA)
Canivez, Gary (Eastern Illinois University, USA)
Guo, Jian-peng (The University of Hong Kong, China)

The purpose of this study was to detect differential item functioning (DIF) in the Adjustment Scales for Children and Adolescents (ASCA) between the US standardization sample (N=709) and the Chinese sample (N=554), for students in Grades 1 – 6. The Simultaneous Item Bias

procedure (SIBTEST) was implemented. Bonferroni adjusted α levels provided correction for maintaining a family-wise α level of .05 through multiple hypothesis tests. Using Roussos and Stout's (1996) criteria (Given p < .05, |βUNI| < .059 refers to negligible DIF; .059 ≤ |βUNI| < .088 refers to moderate DIF; |βUNI| ≥ 0.088 refers to large DIF), preliminary results showed that the Attention-Deficit Hyperactive scale has the most DIF items (7 large DIF items) and no DIF items existed in the Solitary Aggressive (Impulsive) scales, controlling for the overall proficiency by each subscale. Items in the Delinquent scale could not be examined due to no usable score cells. Overall, in the ASCA scales, 8 large DIF items were found against the China sample and 10 large and 2 moderate DIF items were found against the US standardization sample. Further analyses will examine DIF items across sex, grade, and age between the US and Chinese groups. Results of this study contribute to (a) greater understanding of using SIBTEST to detect DIF in a psychopathology measure and possible sources in terms of the sex, developmental, and cultural perspectives and (b) providing psychometric evidence in development of the Chinese version of the ASCA.

Paper 3

"Adjustment Scales for Children and Adolescents factorial invariance: Chinese and U.S. standardization sample comparisons"

Canivez, Gary (Eastern Illinois University, USA)
Kuo, Yi-Lung (University of Iowa, USA)*
Ding, Yi (University of Toledo, USA)
Guo, Jian-peng (The University of Hong Kong, China)

Exploratory factor analysis of the core syndrome raw scores from the Chinese translation of the ASCA using multiple factor extraction criteria, oblique and orthogonal rotation, and coefficients of congruence are reported. Two factors were extracted through principal axis factor analysis based on results from four of six factor extraction selection criteria (eigenvalues > 1, the scree test, HPA, and theoretical consideration). MAP and standard error of scree analyses indicated that only one factor should be extracted. Results of oblique rotation (Promax) for the two factors extracted indicated the ADH, SAP, SAI, and OPD core syndromes were strongly associated with the first factor (Overactivity) while the DIF and AVO core syndromes were strongly associated with the second factor (Underactivity). The correlation between Factor 1 (Overactivity) and Factor 2 (Underactivity) based on the promax rotation was .23, supporting the independence of the Overactivity and Underactivity dimensions and viability of an orthogonal solution. Orthogonal (Varimax) rotation of the two factors also resulted in the ADH, SAP, SAI, and OPD core syndromes having strong associations with the first factor (Overactivity) while the DIF and AVO core syndromes had strong associations with the second factor (Underactivity). Coefficients of congruence (Watkins, 2005) are presented, as are salient variable similarity indexes, and tested the factorial invariance of the present factor structure results to the total ASCA standardization sample (McDermott, 1993, 1994). Results were generally an "excellent" or "good" (MacCallum, et al., 1999, p. 93) match to the factorial results from the U.S. ASCA standardization sample.

Paper 4

"Adjustment Scales for Children and Adolescents item prevalence and mean scale differences: Chinese and U.S. standardization sample comparisons"

Canivez, Gary (Eastern Illinois University, USA)
Ding, Yi (University of Toledo, USA)*
Kuo, Yi-Lung (University of Iowa, USA)
Guo, Jian-peng (The University of Hong Kong, China)

This paper describes several comparisons of the Chinese sample to the U.S. standardization sample. First, examination of the prevalence/base rates of positive behaviors found that 24 of the 26 positive ASCA behaviors were endorsed in ≥ 50% of the Chinese sample as observed in the U.S. sample. Of the 24, 17 showed no significant differences in prevalence/base rates between the Chinese and U.S. samples. Comparisons between the Chinese sample and U.S. standardization sample with respect to the 20 most common problem behaviors and the 20 most rare problem behaviors showed similarities in specific behaviors as well as prevalence estimates. As observed in the U.S. sample, the overwhelming majority of rare problem behaviors were related to overactive, aggressive, and delinquent behaviors. ASCA core syndrome, supplementary syndrome, and overall adjustment scale raw scores from the Chinese sample (N = 554) were compared to the ASCA standardization sample (N = 709) using MANOVA and ANOVA. Both MANOVAs were statistically significant but univariate ANOVAs for ASCA core syndrome, supplementary syndrome, and overall adjustment scale raw scores showed only 2 statistically significant differences that were due to the large sample sizes. All effect sizes (η2 and d) were judged trivial and not clinically meaningful, falling well short of the minimum values for even a small effect size (Cohen, 1988). Results indicated remarkably similar Chinese teacher ratings of Chinese children on the Chinese ASCA as U.S. teacher ratings of U.S. children on all scales of the ASCA.

(ii)Development and psychometric assessment of a learning behaviors measure for use in Mainland China: Learning Behaviors Scale (Chinese Mandarin Translation)

Chair

Ding, Yi   (University of Toledo, USA)

Symposium Abstract

This symposium presents a collection of papers that detail the development and psychometric evaluation of a Chinese Mandarin translation of the Learning Behaviors Scale (LBS; McDermott, Green, Francis, & Stott, 1999) for use in mainland China. The Learning Behaviors Scale is a teacher-report questionnaire designed and found to measure student behaviors related to effective and efficient learning and normed on a representative U.S. sample (McDermott, 1999). By identifying deficit learning behaviors that impact student learning, and which are amenable to intervention, teachers may improve learning outcomes when coupled with effective instruction. A search for existing Chinese instruments measuring learning behaviors or academic enablers revealed that such instruments have yet to be developed. Paper 1 details the methods and procedures of translation, back-translation, and consideration of cultural relevancy of LBS items and dimensions, and item modification where needed; in addition to data collection procedures that produced a large sample of Chinese children (N = 554) in first through sixth grade (ages 6-13). Paper 2 concerns differential item functioning comparisons between the Chinese and comparably aged and grade placed children from the U.S. LBS standardization sample as well as across sex and development. Paper 3 examines the factorial invariance of the LBS to determine if the items measured similar dimensions for the Chinese sample compared to the LBS standardization sample. Paper 4 details examination of subscale (Competence/Motivation, Attitudes Toward Learning, Attention/Persistence, Strategy/Flexibility) and Total scale differences between the Chinese children and children in the LBS standardization sample.

Paper 1

"Translating the Learning Behaviors Scale into Chinese Mandarin and field testing data collection in Mainland China"

Ding, Yi (University of Toledo, USA)*
Kuo, Yi-Lung (University of Iowa, USA)
Canivez, Gary (Eastern Illinois University, USA)
Guo, Jian-peng (The University of Hong Kong, China)

Learning behaviors have been viewed as keystone or learning-to-learn skills essential to school success because of their relative mutability and their correlations to academic achievement. Basic learning behaviors are teachable and are indicatives of a substantial portion of the variability in school success. The LBS (McDermott, Green, Francis, & Stott, 1999) is a 29-item observation device completed by a student's classroom teacher along dimensions of student competence/motivation, attention/persistence, strategy/flexibility, and attitudes toward learning. This paper presents the design, translation, and data collection for the Learning Behaviors Scale (LBS) in an independent sample of 554 Chinese students ranging from first to sixth grade in an inner city elementary school. The translation of LBS to the Chinese Mandarin version included forward translation, backward translation, comparison, revision where necessary, and field testing. The Chinese LBS was examined item-by-item to identify the cultural relevance of the item stems and behaviors from a Chinese perspective. Descriptive data of the sampling, demographic information, and the LBS are presented. Issues involved in cross-cultural assessment, especially the importation of an instrument from a Western culture, are addressed with respect to item bias, interpretation bias, and challenges associated with linguistic, concept, and metric equivalence.

Paper 2

"Learning Behaviors Scale differential item functioning: Chinese and U.S. standardization sample comparisons"

Kuo, Yi-Lung (University of Iowa, USA)*
Ding, Yi (University of Toledo, USA)
Canivez, Gary (Eastern Illinois University, USA)
Guo, Jian-peng (The University of Hong Kong, China)

The purpose of this study was to examine differential item functioning (DIF) in the Learning Behaviors Scale (LBS) between the US standardization sample (N=709) and the Chinese sample (N=554) in Grades 1 – 6. Given the likert-type scores, the Polytomously Simultaneous Item Bias procedure (POLY-SIBTEST) was implemented. Bonferroni adjusted α levels provided correction for maintaining a family-wise α level of .05 through multiple hypothesis tests. Using Roussos and Stout's (1996) criteria (Given $p < .05$, $|\beta UNI| < .059$ refers to negligible DIF; $.059 \leq |\beta UNI| < .088$ refers to moderate DIF; $|\beta UNI| \geq 0.088$ refers to large DIF), preliminary results show that the Strategy/Flexibility scale has the most DIF items (3 large and 1 moderate DIF items), and the Competence/Motivation scale has the least DIF items (2 large items), controlling for the overall proficiency by each subscale. Items in the Attitudes Toward Learning scale could not be examined due to the only one item in the matching subtest. Overall, in the LBS scales, 6 DIF items were found against the China sample and 3 DIF items were found against the US standardization sample. Further analyses will examine DIF items across sex, grade, and age between the US and Chinese groups. Results of this study contribute to (a) greater understanding of using POLY-SIBTEST to detect DIF in behavioral assessment and possible sources in terms of the sex, developmental, and cultural perspectives and (b) providing psychometric evidence in development of the Chinese version of the LBS.

Paper 3

"Learning Behaviors Scale mean scale differences: Chinese and U.S. standardization sample comparisons"

Canivez, Gary (Eastern Illinois University)
Ding, Yi (University of Toledo)*
Kuo, Yi-Lung (University of Iowa)
Guo, Jian-peng (The University of Hong Kong, Hong Kong SAR, China)
Yang, Ling-yan (The University of Iowa, USA)
McDermott, Paul A. (The University of Pennsylvania, USA)

This paper describes several comparisons of LBS subscale scores and the Total score between the Chinese sample and the U.S. standardization sample. LBS subscales and total raw scores from the present Chinese sample (N = 554) were compared to the LBS standardization sample (N = 751) using MANOVA and ANOVA. ANOVA analyses α levels were adjusted with Bonferroni correction for multiple significance tests. The MANOVA for LBS subscale scores (CM, AL, AP, SF) comparing the Chinese sample and subjects within the LBS standardization sample was statistically significant: Wilks $\Lambda = .86$, $F(4, 1300) = 51.84$, $p = .0001$, partial $\eta 2 = .138$. Univariate ANOVAs for the LBS subscales were all statistically significant. ANOVA for the LBS Total raw score was also statistically significant, $F(1, 1300) = 129.34$, $p < .0001$, partial $\eta 2 = .090$. Statistical significance may be primarily related to the large sample sizes but other factors may also play a role. Descriptive statistics for the Chinese and LBS standardization sample showed the Chinese students were rated somewhat lower on the LBS subscales and LBS total raw scores with effect sizes that were small (CM $d = .409$, AP $d = .315$, SF $d = .470$) to medium (AL $d = .664$, Total $d = .636$) (Cohen, 1988).

Paper 4

"Learning Behaviors Scale factorial invariance: Chinese and U.S. standardization sample comparisons"

Canivez, Gary (Eastern Illinois University, USA)
Ding, Yi (University of Toledo, USA)
Kuo, Yi-Lung (University of Iowa, USA)*
Guo, Jian-peng (The University of Hong Kong, Hong Kong SAR, China)
Yang, Ling-yan (The University of Iowa, USA)
McDermott, Paul A. (The University of Pennsylvania, USA)

Exploratory factor analysis of the item raw scores from the Chinese translation of the LBS using multiple factor extraction criteria, oblique and orthogonal rotation, and coefficients of congruence are reported. While the eigenvalues > 1 criterion suggested extracting five factors, the other four criteria (scree test, standard error of scree, HPA, and theoretical consideration) suggested four factors. Four factors were extracted through principal axis factor analysis with equamax rotation for comparison to the LBS standardization sample using congruence coefficients (Watkins, 2005) that ranged from .85 to .93. While these coefficients suggest fair to good match, significant problems were noted in few items obtaining salient factor loadings for two of the factors (Factor III and IV) and item migration problems were observed throughout. A separate EFA with all 29 items from the Chinese LBS was conducted and produced a two-factor model that satisfied three primary criteria also used by McDermott (1999): (a) Cattell's scree test, (b) having a minimum of 5 items with salient factor loadings (≥ .40), and (c) adequate internal consistency estimates for dimensions based on salient items (α ≥ .70). Nine of the 29 Chinese LBS items failed to load on either of the two factors but unlike the U.S. version of the LBS, the Chinese version did not have any items that cross-loaded (salient loadings on multiple factors) so item assignment to factors is unique. Implications for the different factor structure of the LBS in the Chinese sample are discussed.

S12 10:30-12:00 LT4

Current issues in assessment: Perspectives from New Zealand

Chair

Gilmore, Alison (University of Canterbury, New Zealand)

Symposium Abstract

The New Zealand government is introducing multiple educational initiatives which have implications for the roles and practices of assessment and testing across sectors in the education system. The most prominent current initiative is the implementation of national literacy and mathematics standards for all primary schools. The symposium is presented by members of the New Zealand Assessment Academy (a group of senior assessment and measurement experts) and discusses four aspects of nationally important issues related to assessment within the New Zealand context.

Paper 1

"National literacy and mathematics standards in New Zealand: Keeping our sights on an educationally sound approach"

Gilmore, Alison (University of Canterbury, New Zealand)*
Smith, Lisa (University of Otago, New Zealand)

In 2010, national standards in writing, reading and mathematics will be implemented in all New Zealand primary schools. The national standards are to be used to monitor (and address) the progress of all children from year 1 to 8; and to enable 'plain reporting' of children's progress to parents. The New Zealand model differs in significant ways from national testing models used in the US, England and Australia which have similarly sought to raise the literacy and numeracy levels of their children. This paper outlines the approach taken to implementing national standards in New Zealand, and discusses the potential for the model to have educational sound outcomes for students and the profession.

Paper 2

"Detailing achievement at the national level in New Zealand"

Smith, Jeffrey K. (Educational Assessment Research Unit, University of Otago, New Zealand)*

New Zealand's new National Standards programme for assessment at years 1-8 calls for determination of which students are at, above, and below "National Standards" in reading, writing, and mathematics for their year in school. The question arises as to just how well New Zealand's students are doing in these areas. What can a year 4 student who is above or below average actually do with regard to writing ability? How much variability is there amongst students at a year level, or amongst students within a given level of performance? This paper examines these questions using data from New Zealand's National Education Monitoring Project and provides examples of innovative graphical approaches to presenting the information. Results are presented for random samples of students in New Zealand at years 4 and 8.

Paper 3

"Online vs offline test performance"

Darr, Charles (New Zealand Council for Educational Research, New Zealand)*
Shih, Paul (New Zealand Council for Educational Research, New Zealand)

Many schools in New Zealand make use of a series of Progressive Achievement Tests (PATs) to monitor student progress. Until recently these tests have been completed in paper and pencil format. For the last three years however, schools have been able to administer some of the tests online. The online version of the test was designed to be comparable to the paper version. This presentation looks at how the test is performing online and compares this with the performance of the test in its pencil and paper format. It also explores students' online response behaviour in terms of their time taken to respond to individual items and patterns related to reviewing and changing answers. Time taken to respond to test items is used to explore the prevalence of guessing on the test.

Paper 4

"Assessing how pre-service teachers learn to become 'assessment capable'"

Smith, Lisa (Univerity of Otago, New Zealand)*
Gilmore, Alison (University oF Otago, New Zealand)
Hill, Mary (University of Auckland, New Zealand)
Cowie, Bronwen (University of Waikato, New Zealand)

This paper presents initial findings from a Teaching and Learning Research Initiative (TLRI) grant in New Zealand. The TLRI is a collaborative effort of Auckland, Canterbury, Otago, and Waikato Universities, focusing on enhancing understanding of how pre-service teachers learn to use assessment in the service of learning. The research questions are (a) what do pre-service teachers know and believe about assessment at entry, part way through and at exit from their teacher education programmes? and (b) in what ways does teacher education, including practica, scaffold pre-service teachers' assessment capabilities? Participants comprise teacher education students from the four universities in their first and third year primary (and at Canterbury, early childhood) programmes. At the time of this presentation, approximately 900 students will have completed a questionnaire designed to explore their beliefs and knowledge about assessment for learning. The development of the questionnaire and the initial results will be presented.

Discussant

"A healthy state of assessment deep down under"

Hattie, John (University of Auckland, New Zealand)*

As well as a critique of the papers, themes and future directions from the research outlined in the papers will be discussed.

S18 10:30-12:00 LT3

Yes you can: Working with in-house counsel

Chair

Bernstine, Daniel O. (President, Law School Admission Council, USA)

Symposium Abstract

This symposium will examine the legal issues that confront testing-organization staff and their lawyers on a day-to-day basis. Included will be a review of common legal issues, including test security, data protection and privacy, protection of intellectual property (trademarks, copyrights, trade secrets, and patents), managing international relationships, accommodations for test takers with disabilities, managing litigation, and international tax and business models. The session will offer the perspectives of two testing agency in-house attorneys, along with those of a psychometrician who has dealt with them on many of these issues. Strategies for a successful relationship will be discussed and shared, and best-practice suggestions from the audience (along with good war stories) will be solicited.

Paper 1

"Standardized testing in a non-standard legal world"

Van Tol, Joan (Law School Admission Council, USA)*

One of the challenges facing standardized testing companies that wish to operate on a global stage is the myriad legal systems and requirements that they face after crossing their own borders. Although it is tempting, for any given issue, simply to apply worldwide the standards of the most stringent country, the reality is not always so simple. These standards often conflict in ways that make the adoption of a single global approach to the resolution of a legal issue impossible. This presentation will review some examples of issues where the adoption of a single, global standard is tempting but not necessarily feasible, including test-taker identification requirements and other security measures; privacy and data protection; and disability accommodations.

Paper 2

"Doing business globally"

Vaseleck, James (Law School Admission Council, USA)*

Countries around the world have vastly different regulatory schemes for foreign corporations wishing to do business, and some are much more open to direct operations than others. Inevitably, a testing company that hopes to do business outside its home country must come to grips with the specific requirements of its target countries, and develop flexible business models that reflect the varying in-country conditions it finds. Doing so can be difficult for testing companies, which are accustomed to maintaining tight controls over their products and services and often resist attempts by outsiders to change anything that they do. Market conditions, cultural norms, and longstanding traditions also can form barriers to international expansion. This presentation will provide examples from LSAC's exploration of overseas opportunities, and will invite participants to share their own experiences and solutions.

Paper 3

"Test security: The roles of psychometricians and lawyers"

Pashley, Peter (Law School Admission Council, USA)*

This presentation will discuss the roles of testing staff and lawyers generally, and will focus specifically on matters of test security. Among the topics discussed will be the preparation of evidence for test cheating cases (how much, and what kind of, evidence is enough?); how much protection for forensic psychometrics is necessary; preparations for, and participation in, litigation; how to spot a legal issue; and when to take an issue to your lawyer. Again, participant input will be invited. This presentation will conclude with tips on making the attorney/testing-staff relationship as successful as possible, from the perspectives of both sides

Discussant

Rudner, Lawrence (Graduate Management Admission Council, USA)*

We had planned to ask the discussant to respond to and add perspectives about the topics covered by the other speakers, without preparing an independent presentation or paper.

S13 13:30-15:00 LT2

Advanced issues in computerized adaptive testing and computerized classification testing

Chair

Wang, Wen Chung (Assessment Research Centre, The Hong Kong Institute of Education, Hong Kong SAR, China)

Symposium Abstract

Computerized adaptive testing (CAT) has been widely implemented in recent years, mainly because of the significant progress in computer technology and item response theory. A major advantage of CAT is that it yields person estimates more efficiently than paper-and-pencil tests. In some cases where a classification of test-takers into a few categories is sufficient, CAT can be adapted to fulfill this goal. The adapted procedure is called computerized classification testing (CCT). Nowadays, many conventional paper-and-pencil tests or non-adaptive tests have been gradually replaced by CAT or CCT. Computerized adaptive testing (CAT) has been widely implemented in recent years, mainly because of the significant progress in computer technology and item response theory. A major advantage of CAT is that it yields person estimates more efficiently than paper-and-pencil tests. In some cases where a classification of test-takers into a few categories is sufficient, CAT can be adapted to fulfill this goal. The adapted procedure is called computerized classification testing (CCT). Nowadays, many conventional paper-and-pencil tests or non-adaptive tests have been gradually replaced by CAT or CCT. The symposium contains four papers which address some advanced issues in CAT and CTT. In the first paper, CCT was implemented under the generalized graded unfolding model. In the second paper, CCT was implemented under the higher-order item response model. In the third paper, CAT was implemented under the two-parameter testlet model with ability-based guessing. In the fourth paper, a new method was developed to increase efficiency in the expected a posteriori ability estimation under multidimensional item response models. All the four papers adopted simulations to evaluate the performances of the new algorithms under a variety of conditions.

Paper 1

"Implementation of computerized classification testing under the generalized graded unfolding model"

Liu, Chen-Wei (Department of Psychology, National Chung Cheng University, Taiwan)
Wang, Wen-Chung (Assessment Research Centre, The Hong Kong Institute of Education, Hong Kong SAR, China)*

The generalized graded unfolding model (GGUM) has been recently developed to describe item responses to Likert items (agree-disagree) in attitude measurement. In this study, we developed two item selection methods in computerized classification testing under the GGUM, the current-estimate/ability-confidence-interval method and the cut-score/sequential-probability-ratio-test method and evaluated their accuracy and efficiency in classification through simulations. The results indicated that both methods were very accurate and efficient. The more point each item had, the fewer the classification categories were, the more accurate and efficient the classification would be. However, the latter method may yield a very low accuracy in dichotomous items with a short maximum test length. Thus, if it is to be used to classify examines with dichotomous items, the maximum text length should be increased.

Paper 2

"Computerized classification testing under the higher-order IRT model"

Lee, Kung-Hsien (Department of Psychology, National Chung Cheng University, Taiwan)*
Wang, Wen-Chung (Assessment Research Centre,The Hong Kong Institute of Education, Hong Kong SAR, China)

Many CCT algorithms have been developed under unidimensional item response theory (IRT) models. In practice, a test battery may contain multiple tests, and a test may contain multiple subtests. For example, an English test usually includes subtests of listening, speaking, reading, and writing, and each subtest consists of several items. Besides, it is theoretically justifiable that there is an upper-layer "overall" English proficiency that governs the four "domain" abilities of listening, speaking, reading, and writing. Under such a case, the higher-order IRT model is needed, because it accommodates such a hierarchical structure in latent traits. The major advantage of the higher-order IRT model is that both the overall and domain abilities can be estimated simultaneously, which is not applicable for standard unidimensional or multidimensional IRT models. Until now, CCT has not yet been developed under the higher-order IRT model. In this study, we developed such algorithms and evaluated their performances under a variety of situations through simulations. The results showed that the developed CCT algorithms were efficient and accurate in the categorization of test-takers by taking into account both the overall and domain abilities, especially when the loadings of the overall ability on the domain abilities were high.

Paper 3

"Computerized adaptive testing under the two-parameter testlet model with ability-based guessing"

Huang, Sheng-Yun (Assessment Research Centre, The Hong Kong Institute of Education, Hong Kong SAR, China)*
Wang, Wen-Chung (Assessment Research Centre, The Hong Kong Institute of Education, Hong Kong SAR, China)

Testlet response models were developed to fit item responses to testlet-based items. As the guessing in multiple-choice often involves ability, a new IRT model with ability-based guessing was proposed. To analyze multiple-choice items with testlet design, in this study we incorporated the modeling of ability-based guessing into the two-parameters testlet response model, implemented CAT algorithms, and compared the performances of three item exposure control methods through a series of simulation. In the first simulation study, three independent variables were manipulated: (a) testlet effect (small and large), (b) α size (small to large proportion of ability on guessing), and (c) testlet length. In the second simulation study, we compared the performances of three item selection procedures. The results indicated that the CAT algorithms and the three item exposure control methods for the new IRT model were successfully developed and implemented. The smaller the testlet effect and the longer the testlets were, the smaller the root mean square error would be; the Sympson and Hetter online method and the Sympson and Hetter online with progression method could maintain a well-controlled item exposure rate as their pre-specified rate without substantial loss in measurement accuracy. Although the progression method could maintain control of item exposure rate, it had a higher bank usage and higher measurement accuracy.

Paper 4

"Improving the expected a posteriori (EAP) ability estimation in multidimensional computerized adaptive testing"

Chen, Po-Hsi (Department of Educational Psychology and Counseling, National Taiwan Normal University, Taiwan)*

In multidimensional computerized adaptive testing (MCAT), the computer time for traditional expected a posteriori (EAP) ability estimation methods increases exponentially as the number of dimensions increases linearly. For example, in four dimensional MCAT, it took about 15 seconds to yield ability estimates on the four dimensions and to select the next item using traditional EAP estimation method when there were 30 quadrature points in each dimension (Chen, 2006). The computer time should be largely reduced to make MCAT feasible. In this study, I proposed a new EAP estimation method and evaluated its performance through simulations. Six sets of simulated data, constructed by three different numbers of dimensions and two different types of correlations between dimensions, were used to compare the performances of the new and traditional EAP estimation methods. The dependent variables were conditional bias and root mean square of error (RMSE) after administering 5 and 10 items in each dimension, and the mean computer time for ability estimation and item selection. The results indicated that the new method needed much less computer time than the standard method, especially when there were as many as 6 dimensions. In addition, the new and standard methods yielded a similar degree of conditional bias and RMSE. In conclusion, the new method can improve the proficiency of the EAP ability estimation to some degree, and can be easily implemented in MCAT.

S14 13:30-15:00 LT3

Controlling test overlap in computerized adaptive testing

Chair

Chen, Shu-Ying (Department of Psychology, National Chung Cheng University, Taiwan)

Symposium Abstract

Test security is an important concern in computerized adaptive testing (CAT). When test takers can easily obtain test information from previous examinees and answer items correctly simply based on the item pre-knowledge rather than on their proficiency, the observed test scores would be invalid. The threat of item sharing among examinees could be reduced by conducting test overlap control. It is clear that high test overlap would cause a serious threat to test security. With a high test overlap rate, examinees would have an increased chance of obtaining test information from previous test takers. These examinees, however, would not benefit much from item sharing if test overlap is tightly controlled and the number of common items among examinees is small. Test overlap commonly considered in CATs is defined as the average proportion of items shared by pairs of examinees and can be effectively controlled by implementing the SHT procedure (Chen & Lei, 2005). Controlling this pairwise test overlap, however, may not be sufficient for dealing with item sharing observed in the field. In practice, examinees may obtain test information from more than one previous test taker. This larger scope of information sharing needs to be considered in conducting test overlap control. The purpose of this symposium is to modify the SHGT procedure proposed by Chen (in press) such that the proportion of common items between an examinee and a group of previous test takers can be well controlled.

Paper 1

"Controlling general test overlap in computerized adaptive testing"

Chen, Shu-Ying (Department of Psychology, National Chung Cheng University, Taiwan)
Chen, Jyun-Hong (Department of Psychology, National Chung Cheng University, Taiwan)*

The purpose of this study is to propose a test overlap control method, the modified Sympson and Hetter procedure with general test overlap control (MSHGT). The MSHGT procedure is designed to control the general test overlap rate and item exposure rates simultaneously on the fly without any iterative simulations conducted prior to operational CATs. To investigate the effect of the MSHGT procedure on test overlap control and measurement precision, results observed from other on-the-fly exposure control procedures were used to evaluate the extent of improvement in test security and loss of ability estimate precision in the

MSHGT procedure. Results indicated that item exposure rate and general test overlap rate can be simultaneously controlled by implementing the MSHGT procedure. In addition, these two indices were controlled on the fly without any iterative simulations conducted prior to operational CATs. Thus, the MSHGT procedure would be an efficient method for controlling item exposure and general test overlap in CATs.

Paper 2

"Controlling general test overlap conditional on ability in computerized adaptive testing"

Lin, Yi-Hung (Department of Psychology, National Chung Cheng University, Taiwan)*
Chen, Shu-Ying (Department of Psychology, National Chung Cheng University, Taiwan)

Even though general test overlap could be well controlled by implementing the MSHGT procedure, it does not guarantee that general test overlap at a given ability level is also well controlled. The general test overlap rate may be close to 1.0 for examinees at a particular ability level, even though that is low for examinees across all ability levels. In order to improve test security by ability levels, controlling general test overlap at ability levels is necessary. To control item exposure and general test overlap conditional on ability levels, the MSHGT procedure could be adapted to the Stocking and Lewis (1998, 2000) conditional procedure. Preliminary results indicated that the general test overlap at middle ability levels were better controlled than those at extreme ability levels. Further studies are currently undertaken to improve general test overlap control for examinees with extreme ability levels.

Paper 3

"Controlling general test overlap conditional on testing time and testing location in computerized adaptive testing"

Liu, Tzu-Chen (Department of Psychology, National Chung Cheng University, Taiwan)*
Chen, Shu-Ying (Department of Psychology, National Chung Cheng University, Taiwan)

In addition to the conditioning on ability level, the general test overlap conditioning on testing time or testing location should also be considered because an individual would more likely get test information from the most recent examinees than from examinees who took the test long time ago or from neighboring examinees than from remote examinees. To control the general test overlap conditional on testing time or testing location, the MSHGT procedure could be applied with item usage counted among the group of examinees who take the test in close proximity or reside in neighboring areas rather than among all examinees who have taken the test. The flexibility of the MSHGT procedure in controlling general test overlap conditioning on testing time and testing location is currently under investigation.

Paper 4

"Improving test security for a-stratified CAT by implementing general test overlap control"

Lee, Hsiang-Ling (Department of Psychology, National Chung Cheng University, Taiwan)*
Chen, Shu-Ying (Department of Psychology, National Chung Cheng University, Taiwan)

The a-stratified design was proposed to equalize uneven item usage by stratifying an item pool based on the a-parameter value. At the early stages of a test, items are administered from the stratum with lower a-parameter values, and those with higher a-parameter values are administered during the later stages. By implementing the a-stratified design, there are few items with exposure rates equal to zero, and item usage is well-balanced without sacrificing the efficiency in ability estimation (Chang & Ying, 1999). Even though the a-stratified design has been shown to be effective in improving measurement efficiency (Chang & Ying, 1999), without tight control on test overlap, the observed general test overlap rate may be too high to be acceptable in practice. The purpose of this study is to improve test security for the a-stratified design by implementing the MSHGT procedure.

S15 13:30-15:00 LT4

A combined emic-etic approach to personality assessment: Recent applications of the CPAI-2 in Chinese settings

Chair

Gan, Yiqun (Peking University, China)

Symposium Abstract

This symposium describes a case study of two companies, ATA and SHL, operating in different verticals and geographies working together to meet a pressing assessment need in China. The project was to provide a short competency-based assessment for identifying the fit of candidates to jobs organised into several job families. The requirements were or a short, efficient but psychometrically sound tool organised around a clear taxonomy of behavioural competencies to compliment a taxonomy of technical competencies already identified for over 600 jobs ranging from managerial and supervisory levels to skilled and semi-skilled operational roles, and covering roles within information technology through to marketing and sales as well as administration. The symposium describes the assessment need, how the assessment technology originally developed in Europe was adapted for the Chinese market, how the solution was deployed and experience since original deployment. Key learnings from the project are also described including how the project has served to show the importance of softer behavioural competencies to companies in the growing Chinese economy.

Paper 1

"Differences in CPAI-2 personality traits among Chinese college students from different study fields"

Ng, Alexander (Department of Psychology, Chinese University of Hong Kong, Hong Kong SAR, China)*
Cheung, Fanny (Department of Psychology, Chinese University of Hong Kong, Hong Kong SAR, China)
Fan, Weiqiao (Department of Psychology, Shanghai Normal University, China)
Leong, Frederick (Department of Psychology, Michigan State University, USA)

Choosing an area of study is an early milestone in one's career life. A clear understanding of college students' personality profiles from different study fields might help high-school students make better choices. In this study, we used MANOVA to examine the differences in four universal CPAI-2 personality scales [i.e. Leadership, Logical vs. Affective Orientation, Aesthetics, and Extraversion vs. Introversion] and four indigenously derived CPAI-2 personality scales [i.e. Face, Self vs. Social Orientation, Renqing/Relationship Orientation, and Social Sensitivity] among 2,236 Hong Kong college students from six different faculties: Arts, Business Administration, Engineering, Medicine, Science, and Social Sciences. The results suggest that, in general, Chinese college students from different faculties differ in the eight CPAI-2 personality traits. To identify the specific patterns of these differences, post-hoc tests were also conducted and the results show that Business students scored significantly higher on the Leadership, Extraversion, Social Orientation, Social Sensitivity, Face, and Renqing scales than Science students, but both of them scored significantly lower on the Aesthetic and Affective Orientation scales than Arts students. More interestingly, similar to Business students, Social Science students also scored significantly higher on the Leadership, Extraversion, and Social Sensitivity scales than Science students, but Social Sciences students score significantly lower on the Renqing scale and score significantly higher on the Aesthetic scale than Business students.

Paper 2

"Teacher leadership behaviors: An indigenous model of leadership effectiveness in the Chinese educational setting"

Yao, Jingdan (Department of Psychology, Chinese University of Hong Kong, Hong Kong SAR, China)*

The study investigated teacher leadership in secondary schools in Southern China with the Behavioral Complexity Model derived from Quinn's (1988) Competing Values Framework. Teachers' self ratings and students' report (e.g., collective identity and school satisfaction) were used to evaluate teacher leadership effectiveness. The Cross-Cultural Personality Assessment Inventory (CPAI-2) was employed to examine the influence of personality on leadership behaviors. A total of 230 head teachers from secondary schools and 10 students of each teacher (N=2300, aged from 13-18 years old) were recruited. The instruments were tested with confirmatory factor analysis (CFA) to determine their applicability in the present sample. Structural equation modeling (SEM) was used to examine the proposed theoretical framework. Implications of the Behavioral Complexity Model and its personality correlates on understanding teacher leadership were discussed.

Paper 3

"The moderating role of harmony in the relationship between proactive personality and contextual performance"

Gan, Yiqun (Department of Psychology, Peking University, China)*

The objective of the present study was to examine the moderating role of Harmony-a CPAI-2 personality scale- in the relationship between proactive personality and job performance. One hundred and fifty-eight employees in Chinese state-owned companies completed Proactive Personality Scale, Harmony scale in CPAI-2, and Job Performance Appraisal Form. Results of the hierarchical regression analyses indicated that when demographic variables were controlled, Harmony had significant moderating effects on job dedication and interpersonal facilitation. In the high Harmony group, the correlation between proactive personality and contextual performance was highly significant; whereas in the low Harmony group, this correlation was not significant. The present results suggest that in China, organizations should also pay attention to the level of Harmony personality during the hiring and selection of candidates.

Paper 4

"Parental influence, personality traits, and readiness for career decision-making among Hong Kong college students"

Fan, Weiqia (Department of Psychology, Shanghai Normal University, China)*
Leong, Frederick (Department of Psychology, Michigan State University, USA)
Cheung, Shu Fai (Department of Psychology, University of Macau, Macau, China)
Cheung, Fanny (Department of Psychology, Chinese University of Hong Kong, Hong Kong SAR, China)

We explored the contributions of perceived family intrusiveness to career readiness with the mediating effect of the personality factor based on the CPAI-2. In Study 1, a paper-and-pencil survey among 392 Hong Kong college students from different study fields indicated that perceived family intrusiveness significantly contributed to career readiness with a significant mediating effect of different personality dimensions. In Study 2, the previous model was further demonstrated with a web-based survey among 2,380 Hong Kong college students. In addition, career decision-making difficulties in readiness showed significant differences among students from different study fields. Students from the fields of business and management, and finance and accountancy showed the lowest level of career difficulties in readiness than students from other fields such as science and humanities. Students from the fields of humanities and arts and design had the highest level of career difficulties in readiness than those students from other study fields. Implications on career development and counseling for college students were discussed.

Discussant

Leung, Kwok (Department of Management, City University of Hong Kong, Hong Kong SAR, China)*

S19 13:30-15:00 LT1

International trends in test security

Chair

Fremer, John (Caveon Test Security, USA)

Symposium Abstract

Organizations involved in international testing have initiated many activities to help protect their examinations and to combat efforts to cheat. Representatives of major organizations in the People's Republic of China, the US and the UK will review the strategies that they have applied and share what they have learned. Issues such as the following will be addressed:

What strategies are being employed? Who has responsibility for these strategies within the testing entities? What seems to be the most effective package of activities? What are the greatest challenges? What is different about protecting against cheating now compared to ten years ago? What have the participants learned about dealing with the media in security cases? What advice can be shared with people just starting out to tackle test security in an international setting?

Paper 1

"International trends in test security - Certification testing"

Fremer, John (Caveon Test Security, USA)*

Certification programs involved in international testing have initiated many activities to help protect their examinations and to combat efforts to cheat. This paper will review the strategies that certification programs have applied and share what they have learned. Issues such as the following will be addressed: What strategies are being employed? Who has responsibility for these strategies within the testing entities? What seems to be the most effective package of activities? What are the greatest challenges? What is different about protecting against cheating now compared to ten years ago? What have the participants learned about dealing with the media in security cases? What advice can be shared with people just starting out to tackle test security in an international setting?

Paper 2

"International trends in test security - Employment testing"

Burke, Eugene (SHL Group, UK)*

Employment testing programs involved in international testing have initiated many activities to help protect their examinations and to combat efforts to cheat. This paper will review the strategies that such programs have applied and share what they have learned. Issues such as the following will be addressed: What strategies are being employed? Who has responsibility for these strategies within the testing entities? What seems to be the most effective package of activities? What are the greatest challenges? What is different about protecting against cheating now compared to ten years ago? What have the participants learned about dealing with the media in security cases? What advice can be shared with people just starting out to tackle test security in an international setting?

Paper 3

"International trends in test security - Testing of English as a second language"

Geranpayeh, Ardeshir (University of Cambridge ESOL Examinations, UK)*

Testing of English as a foreign language programs involved in international testing have initiated many activities to help protect their examinations and to combat efforts to cheat. This paper will review the strategies that such programs have applied and share what they have learned. Issues such as the following will be addressed: What strategies are being employed? Who has responsibility for these strategies within the testing entities? What seems to be the most effective package of activities? What are the greatest challenges? What is different about protecting against cheating now compared to ten years ago? What have the participants learned about dealing with the media in security cases? What advice can be shared with people just starting out to tackle test security in an international setting?

Paper 4

"International trends in test security - National level and large scale testing"

Tong, Alex (ATA, China)*

National and large-scale testing programs involved in international testing have initiated many activities to help protect their examinations and to combat efforts to cheat. This paper will review the strategies that such programs have applied and share what they have learned. Issues such as the following will be addressed: What strategies are being employed? Who has responsibility for these strategies within the testing entities? What seems to be the most effective package of activities? What are the greatest challenges? What is different about protecting against cheating now compared to ten years ago? What have the participants learned about dealing with the media in security cases? What advice can be shared with people just starting out to tackle test security in an international setting?

Discussant

Sun, James Jian-Min (Renmin University of China, China)*

S16 15:15-16:45 LT2

Scale development using Rasch models

Chair

Mok, Magdalena Mo Ching (Assessment Research Centre, & Psychological Studies Department, HKIEd, Hong Kong SAR, China)

Wang, Wen Chung (Assessment Research Centre, & Psychological Studies Department, HKIEd, Hong Kong SAR, China)

Symposium Abstract

This symposium focuses attention on the development of measurement scales using Rasch measurement models. Five papers with authors from five universities from China, Taiwan and Hong Kong are included in the symposium. These papers share in common the advancement of new knowledge in using Rasch models for the development of measurement scales. The paper by Wang, Zheng, and Wang, and the paper by Cheng and Lam are concerned with the development of psychological scales, namely Emotional Intelligence Scale and General Self-Efficacy Scale in the first paper by Wang and associates, and Teacher Motivation Scale in Project Learning in the second paper by Cheng and Lam. The other three papers are concerned with theoretical issues in the development of scales for measuring academic outcomes. Both the paper by Lee, and the paper by Mok, Yan and Lau focus attention on measurement issues in the building of vertical scales across grade levels, including item calibration, test invariance, linkage methods, and sample size considerations. The paper by Tam, Mok, Lau and Wu is an exploration of using user-defined fit statistic to analyse two-tier items in the context of mathematics assessment. Analyses of the three theoretical papers are based on large scale empirical data as well as simulated data. All papers in the symposium are innovative and expected to contribution to new knowledge of testing.

Paper 1

"Rasch model analysis of the Emotional Intelligence Scale and the General Self-Efficacy Scale"

Wang, Li-Jun (Zhejiang Normal University & The Hong Kong Institute of Education, Hong Kong SAR, China)*
Zheng, Xian-Liang (Shanghai Normal University, China)
Wang, Wen-Chung (Assessment Research Centre, & Psychological Studies Department, HKIEd, Hong Kong SAR, China)
Mok, Magdalena Mo Ching (Assessment Research Centre, & Psychological Studies Department, HKIEd, Hong Kong SAR, China)

Although support has been found for the psychometric properties of the Emotional Intelligence Scale (EIS) and the General Self-Efficacy Scale (GSES) using classical test theory approaches, these two scales have not yet been analyzed with Rasch models.The aims of this study were to use Rasch analysis to assess the psychometric properties of the EIS and the GSES, to investigate the relationship between them, and to examine group differences in emotional intelligence and self-efficacy. The Chinese-version of the EIS and GSES were administered to 299 college students. The Rasch partial credit model was fit to the data using the ConQuest software. Differential item functioning (DIF) was assessed. The correlation between these two latent traits was investigated. Latent regression was performed to examine group difference. The revised scales show a good model-data fit and a high reliability. A two-dimensional Rasch partial credit model was fit in order to yield a more accurate estimate of the correlation between the two scales. The correlation was .53, suggesting a moderate relationship between them. The latent regression revealed that family atmosphere and teacher-pupil relationship had a significantly positive regression weight on emotional intelligence, and that gender, family atmosphere, mother's education level, teacher-pupil relationship, student cadre had a significantly positive regression weight on self-efficacy. Rasch analysis is powerful in assessing psychometric properties of a scale, revealing current validity, and examining group difference. It is recommended that Rasch analysis be routinely applied in test development.

Paper 2

"Using user-defined fit statistic to analyze two-tier items in mathematics"

Tam, Hak-Ping (National Taiwan Normal University, Taiwan)*
Mok, Magdalena Mo Ching (Assessment Research Centre, & Psychological Studies Department, HKIEd, Hong Kong SAR, China)
Lau, Doris Ching Heung (Assessment Research Centre, HKIEd, Hong Kong SAR, China)
Wu, Margaret (Assessment Research Centre, University of Melbourne, Australia)

The two-tier item format is relatively new and is gradually gaining popularity in some areas of educational research. A two-tier item is made up of two portions, with the first portion assesses whether students could identify the correct mathematical concept with respect to the information stated in the item stem, while the second portion examines the reason they supplied to justify the concept they chose. Since the data thus collected are related in a certain way, they pose challenges regarding how analysis should be done to capture the relationship that exist between the two tiers. The present paper attempts to analyze such data by using a user-defined fit statistics within the Rasch approach. The kind of information that can be gathered will be illustrated by way of analyzing a data set in mathematics.

Paper 3

"Development and validation of Teacher Motivation Scale in project learning"

Cheng, Rebecca Wing-Yi (Assessment Research Centre, & Psychological Studies Department, HKIEd, Hong Kong SAR, China)*
Lam, Shui-fong (The University of Hong Kong, Hong Kong SAR, China)

According to self-determination theory (Ryan & Deci, 2000), different types of motivation can be placed on a continuum according to the extent they reveal self-determination. From the least self-determined to the most self-determined motivation are (a) external regulation (doing a task for external monitoring), (b) introjected regulation (doing a task for approval from others), (c) identified regulation (doing a task for its importance), and (d) intrinsic regulation (doing a task for enjoyment or interest). Based on this theoretical framework, we developed a scale to measure teachers' motivation in implementing project learning activity. The scale consisted of 20 items grouped in 4 subscales (i.e., external, introjected, identified and intrinsic). Confirmatory factor analysis on data from 182 Chinese teachers from eight secondary schools in Hong Kong supported the four-factor structure of the scale. Results of 1-dimensional Rasch analysis using the Winsteps programme suggested that the response categories functioned well and there was more than one dimension to the data. When the data were subjected to a 4-dimensional Rasch analysis using the Conquest programme, it was found that the data fitted the model well. Overall, the teacher motivation scale was found to be reliable and valid. This instrument provides important resources for the schools that implement project learning activity.

Paper 4

"Development of a vertical EFL ability scale for Hong Kong – Calibration and test invariance issues"

Lee, Tony (Assessment Research Centre, HKIEd, Hong Kong SAR, China)*

The use of linking items / tests in developing vertical ability scales is well known and accepted. In practical scale building, however, the choice of linking items / tests and the anchor values to be used are not as straightforward as it is sometimes assumed. This paper described issues encountered in developing an English as a foreign language (EFL) ability scale for Hong Kong and the solutions adopted. The paper touches on issues relating to customary design approaches in reading comprehension test design for English language. Sub-samples sharing the same linking items pose problems for decision on which items to use for linking and the anchor values to be adopted. The solutions adopted in this paper are able to deliver a vertical ability scale of English with sufficient stability and test invariability for general application.

Paper 5

"Considerations in developing vertical scales using IRT: Link methods, link items and sample size issues"

Mok, Magdalena Mo Ching (Assessment Research Centre, & Psychological Studies Department, HKIEd, Hong Kong SAR, China)*
Yan, Zi (The Hong Kong Institute of Education, Hong Kong SAR, China)
Lau, Doris Ching Heung (Assessment Research Centre, HKIE, Hong Kong SAR, China)

The purpose of this paper is to contribute to recommendations for the development of vertical scales using IRT methods. The recommendations will be based on three sources, namely, test development literature, simulated data, and empirical data. Considerations in vertical scale development include in this study include method (concurrent, stepwise-chain) of linking, properties of link items (goodness of fit, adherence to curriculum), number and proportion of link items (15%, 30%), and model of analysis (Rasch, 2-parameter model). A number of data sets will be simulated to test effect on scale properties of different methods in the construction of vertical scale. Merits of these methods are compared by comparing the match between parameters of the estimated and the original (generated) samples. The real data set comprise a sample of 5,755 primary students between primary 2 and primary 6 from 24 schools, and 3,621 secondary students between secondary 1 and secondary 3 from 11 schools in Hong Kong. The mathematics competencies of participants were assessed using booklets with linked items between adjacent year levels. Different methods of constructing the vertical were compared in terms of estimated population mean (grade-to-grade growth), estimated population standard deviation (grade-to-grade variability), separation of grade distributions by effect size, and differential item functioning. The paper will combine findings from the literature review, analysis on simulated and real data to derive a set of recommendations regarding the development of vertical scales for charting growth across year levels.

S17 15:15-16:45 LT6

Issues of policy, ethics, professionalism and training in multinational testing

Chair
Bartram, Dave (SHL Group Ltd, UK)

Symposium Abstract
This symposium will provide an overview of major developments in the field of international standards and guidelines in tests and testing that have taken place over the past decade. It will focus in particular on the development of professionalism in the use of the tests in work and organizational settings and will include a presentation that focuses on the relevance for China of the issues such guidelines and standards cover.

Paper 1

"The role and function of international guidelines and standards in testing and test use"

Bartram, Dave (SHL Group Ltd, UK)*

The International Test Commission (ITC) and European Federation of Psychologists' Associations (EFPA) have both worked to develop international guidelines and standards for tests and test use. The main ITC developments have been of guidelines on test adaptation, guidelines on test use and guidelines on computer-based testing. The guidelines on test adaptation are currently under review and we expect a new version to be produced in the near future. The ITC guidelines have been very influential in their effect on practice and on the development of professional standards at national level. The Test Use Guidelines, for example, have been adopted by a large number of professional psychological associations and translated into many different languages. They have also been used as the basis for the development of the more detailed EFPA Standards for Test User Certification. In addition to developing a model for Test User Certification and for accrediting national test user certification schemes, EFPA has produced a set of criteria for the review of tests. These are used in a number of countries as the basis for test registration, test certification or test review.

Paper 2

"Survey on the use of tests by psychologists in Europe: 1999 to 2009"

Evers, Arne (University of Amsterdam, Netherlands)*
Muniz, Jose (University of Oviedo, Spain)
Bartram, Dave (SHL Group Ltd, UK)

Various institutions, both national and international, develop projects and activities aimed at improving the quality of tests and test use. In order to make the right decisions in this respect, it is essential to know the opinion of psychologists about tests and testing practices. In 1999 the Committee on Tests and Testing of the European Federation of Psychologists' Associations (EFPA) undertook a survey on attitude towards tests and testing among psychologists in six European countries. It was concluded then, that European psychologists show a positive attitude, although a need for a more active role of institutions was expressed. In 2009 this survey was repeated in 17 European countries (including the six countries of the first survey) with a total of about 12500 respondents. The questionnaire was extended with items concerning computerized tests and testing by Internet. In this presentation the results of the two surveys will be compared and the differences among the 17 countries in the second survey will be discussed. Some possible actions for improving testing practices in Europe will be suggested.

Paper 3

"ISO 10667: An international standard for assessment in work and organizational settings"

Born, Marise (Erasmus University Rotterdam, Netherlands)*

The International Standards Organization project ISO-PC230 was launched under the auspices of the Deutsches Institut für Normung in Berlin in March 2007 with the aim to develop a standard for assessment of people in work settings. The standard's objective is to improve work-related assessment quality services. The standard has a process-oriented perspective implying accountability for the whole assessment service delivery. Evidence based assessment is seen as key to assessment quality. More than 30 countries have become actively involved in the development of this standard, which is aimed to be operational before 2011. This presentation will describe the cycles of development of this ambitious international project and how it may operate as a general standard to which other guidelines for assessment quality may be linked, such as test and test user quality guidelines.

Paper 4

"The application of psychological tests in the field of business in China: Situation and challenges"

Sun, James Jian-Min (Renmin University of China, China)*

Psychological tests were originally used in the education field mainly for scientific research by either scholars or well-trained teachers in pedagogy or psychology in China. With the development of the economy, how to find and select qualified talent became one of the top challenges for executives and human resource professionals in the business field. This, in turn facilitated the wide recognition and acceptance of the usefulness of psychological tests. It is estimated that there are more than 1,500 test service providers in China. These can be divided into four types: 1) Government-sponsored agencies including different types of Human Resource Service Agency on provincial and municipal level throughout the country; 2) Academic institutions affiliated to consulting companies; 3) Domestic business consulting companies; and 4) International consulting companies. The issues we are facing include: 1. Abuse of tests; 2. No standard for test users; 3. No professional conduct ethics; 4. The lack of authority in implementing professional ethics and certification standards in the whole country; 5. The lack of procedures and methods in adaptation and cross-cultural validation of Western tests; 6. The pressure created by the huge demand for tests from the business field. Some suggestions will be discussed including the coordination among international and domestic psychologists and professional associations, the adoption of certification or standards for test users from international associations, such as the ITC and EFPA.

S20 16:45-18:15 LT4

Cultural issues in psychometric assessment and the application of personality measures for organizational use in Hong Kong and China

Chair

Cowleson, Neil (Division of Industrial Organizational Psychology of the Hong Kong Psychological Society, Hong Kong SAR, China)

Symposium Abstract

A highly experienced panel of psychometric experts and test users will share experiences using psychometrics in China, with a specific focus on the cultural issues of using personality measures. Mr U and Mr Fraccaro will share from their corporate perspectives on the application of psychometrics for recruitment and development and the challenges faced by organizations in using such instruments as part of their resourcing and talent management processes in Hong Kong and China. Ms Fung and Ms To, will share their experiences of developing, researching and using two of the world's most widely used personality measures in China. Helen will share findings from over 30,000 administrations of the SHL OPQ and will particularly focus on the construct equivalence between the Chinese samples and a UK English reference sample. Helen will also share analysis of the invariance of between-scale correlation patterns and variations in scale means as a function of language and other demographics. Clara will share findings from the application of the Hogan personality instrument. This will include experiences of cultural differences and issues relating to the use of the Hogan tool in different cultures, with a particular focus on Hong Kong and China. Validity findings for the Hogan tool will be presented and comparisons between different norm samples across different cultures including the Chinese norm. Both Helen and Clara will discuss implications for the interpretation of personality measures in China, and implications for the use of psychometrics in the assessment and development of leaders and other employee/candidate groups.

Paper 1

"Application of psychometrics for recruitment and development"

Fraccaro, Michael (HSBC, China)*

Mr Fraccaro will share from HSBC's perspectives on the application of psychometrics for recruitment and development and the challenges faced by organizations in using such instruments as part of their resourcing and talent management processes in Hong Kong and China

Paper 2

"Application of psychometrics for recruitment and development"

U, Kin Chong (KPMG / DIOP, China)*

Mr. U will share KPMG's experiences on the application of psychometrics for recruitment and development and the challenges faced by organizations in using such instruments as part of their resourcing and talent management processes in Hong Kong and China

Paper 3

"Development, research and the use of personality measures in China"
Fung, Helen (SHL, China)*

Findings from over 30,000 administrations of the SHL OPQ will be reported. In particular, the results focus on the construct equivalence between the Chinese samples and a UK English reference sample. The presentation will report analysis of the invariance of between-scale correlation patterns and variations in scale means as a function of language and other demographics. There will also be a discussion on implications for the interpretation of personality measures in China, as well as implications for the use of psychometrics in the assessment and development of leaders and other employee / candidate groups.

Paper 4

"Development, research and the use of personality measures in China"
To, Clara (Mobley Group Pacific Limited (HK), Hong Kong SAR, China)*

Findings from the application of the Hogan personality instrument will be discussed. This will include experiences of cultural differences and issues relating to the use of the Hogan tool in different cultures, with a particular focus on Hong Kong and China. Validity findings for the Hogan tool will be presented and comparisons between different norm samples across different cultures including the Chinese norm. Implications for the interpretation of personality measures in China, as well as implications for the use of psychometrics in the assessment and development of leaders and other employee / candidate groups will also be discussed.

S21 08:30-10:00 LT2

Executive assessments across cultures

Chair
To, Clara (Mobley Group Pacific Limited (HK), Hong Kong SAR, China)

Symposium Abstract

Facing a complex and ever-changing business environment, organizations have increasingly realized the urgency to ensure business sustainability and enhance its adaptability by identifying and developing the right talents. Organizations attempt to apply various psychological assessment tools to assess and develop talents locally and globally. However, there is lacking concerted efforts to evaluate the impact of culture on different assessment approaches. This symposium presents several practitioners and researchers who have extensively worked on this issue in the Asia Pacific region, Africa, Europe, and the United States. Issues involving the constructs, measurement (and their cross-cultural equivalence), and validities will be addressed. Additionally, implications for executive assessment and development in diverse cultures will be discussed. This symposium should be of great interest to anyone attempting to deploy executive assessment and development approaches in the different cultural contexts.

Paper 1

"Construct equivalence across and between countries: Using forced-choice to control for cultural bias"
Bartram, Dave (SHL Group, Surrey)*

Multinational organizations need to carry out assessments for recruitment and selection or for internal purposes that enable them to make comparisons between people from varying cultural, national and linguistic background. In doing so, they need to use instruments that are construct invariant across these demographic variations. OPQ32 (SHL, 1999, 2006) measures 32 traits that were identified as being of relevance within the world of work. It does so using items that are relatively transparent and work-related. To control for impression management in high stakes assessments, OPQ32i uses a forced-choice 'quad' item format. This can now be scored using a multidimensional IRT model from which latent trait scale scores can be recovered that have normative rather than ipsative properties. With the data available on demographics, it is possible to consider same-language and different-language between-country effects and within country effects relating to cultural factors indexed by first-language and ethnic group differences. Results of analysis from over 74,000 people are reviewed in terms of differences between 19 countries involving 14 different languages, and from a further 50,000 people on effects of between ethnic groups and between first-language groups within one country (South Africa).

Paper 2

"The MLQ transformational – transactional scale and cross-cultural assessment: A research-driven platform for global leadership development"

Elliott, Ray (MLQ International, Australia)*

Thirty-five years of independent blind peer-reviewed research has now followed Bass's seminal publication concerning the Full Range Leadership Model (FRLM) as assessed by the original Multifactor Leadership Questionnaire (MLQ). Many international studies in diverse cultures and organizational contexts have contributed to the acceptance of this transformational – transactional leadership paradigm and to the validation of the shorter 45 item MLQ5x scale. Examining extensive data-sets in global MLQ applications Antonakis, Avolio, and Sivasubramaniam (2003) concluded that the original nine factor model proposed by Bass was supported across data-sets when local context variables were controlled for as moderators. Consequently the MLQ5x occupies a respected position in transparent critical leadership research and it is frequently used as the research benchmark for leadership. Consequently, practitioners concerned with the assessment and development of leaders in cross-cultural contexts find the research-defined outcome orientation of the International MLQ Report important for developing optimal behavioral profiles predictive of many desired inter-personal and objective organizational outcomes. Cross-cultural comparisons using the MLQ5x scale are highlighted and important issues relating to 360 feedback in some cultures are considered.

Paper 3

"Are context-aligned management practices applicable across nations? An investigation of Behavioral Observation Scales"

Sue-Chan, Christina (City University of Hong Kong, Hong Kong SAR, China)*
Li, Stan X (York University (Canada), Canada)
Yao, Xiaotao (Xi'an Jiaotong University, China)

The measurement of performance in work and other performance contexts (e.g., education, sports), has been the subject of much research in North America since the 1950s. North American scholars acknowledge that performance is "complex, dynamic, and multidimensional" (Hough & Oswald, 2000, p. 633); yet, there is a relative paucity of research that has examined the measurement and nature of performance in a context outside of the Western cultural context. We argue that one performance assessment method with empirically demonstrated validity and reliability (e.g., Sue-Chan & Latham, 2004) that is widely used in North America, behavioral observation scales (BOS) (Latham & Wexley, 1994) is applicable across cultures. This is because BOS fall into the category of what we refer to as "context-aligned" management practices. Unlike context-negated practices that are adopted in settings without accounting for contextual factors (e.g., GATB), context-aligned practices are applicable across different contexts but the reason for their predictive validity is influenced by culture values. Our empirical examination showed that BOS are valid in both China and Canada, and their validity stem from the atypical behaviors that are defined by the cultural context in which BOS are developed. Implications for adopting context-aligned measurement and management practices are offered.

Paper 4

"Validity of assessment centre in executive selection and development: A tale of Chinese executives"

To, Clara (Mobley Group Pacific (Hong Kong), Hong Kong SAR, China)*

The validity of assessment centers has been debated ever since they became a popular means for measuring work related behaviors. As this approach has become ever more popular in the past decade for selecting executives worldwide, the need to pursue efforts to determine its integrity has become ever more pressing. The present study attempts to investigate the validity of an assessment center by examining the Profile of Success (POS) for leadership roles derived from different companies electing to employ an assessment center method. Common to POS is the tendency to address adaptability to change; learning agility; drive toward results; ability to work in the face of ambiguity; effective communication. We will conduct a qualitative analysis to study narrative comments that were developed in the form of psychological reports and derived from the assessment center which was aimed at measuring these POS constructs. These narrative comments are based on a participant's career history, personality metrics, cognitive metrics, interview data and behavioral observations obtained through simulation evaluations. Finally, resulting recommendations to the organizations will be examined. Issues relating to the constructs of POS and their proposed measurements, as well as implications of leadership effectiveness among Chinese executives will be discussed.

Discussant

Leung, Kwok (Department of Management, City University of Hong Kong, Hong Kong SAR, China)*

S22 08:30-10:00 LT3

Beyond Chinese culture: The application of Cross-Cultural Personality Assessment Inventory-2 (CPAI-2) in Western and Eastern cultures

Chair

Cheung, Shu Fai (University of Macau, Macau, China)
Born, Marise (Erasmus Unviersity, Rotterdam, Netherlands)

Symposium Abstract

The Cross-Cultural Personality Assessment Inventory-2 (CPAI-2) was developed initially as an indigenous and comprehensive instrument to assess personality in Chinese societies. Combined emic-etic approach was adopted to include traits highly relevant to Chinese societies but not emphasized in prevalent Western models of personality. Previous studies reveal a four-factor structure of personality traits, in which the interpersonal relatedness factor was not covered in the five-factor model. To enrich our understanding of personality traits, it is necessary to administer the indigenous inventory to non-Chinese samples, both Eastern and Western, to examine the structure and external correlates of the traits. In the symposium, studies in diverse cultural groups will be presented, including data from Dutch, Vietnam, Romania, China, South Korean, Japan, and North America. Born will examine the added value of the CPAI-2 above the HEXACO-PI-R in predicting student-related criteria in Dutch. Iliescu and Ion will discuss the incremental validity of the CPAI-2 traits in predicting leadership behaviors in Romanian. Leong, Fan, and colleagues will examine the utility of CPAI-2 in predicting career exploration among university students in US and Hong Kong. Dang, Weiss, and colleagues will illustrate the adaption of the CPAI-A (the adolescent version of the CPAI-2), in Vietnam. S.Cheung and F.Cheung will examine the relationships between CPAI-2 traits and correlates related to openness and interpersonal relatedness among Asian and American college students. This symposium illustrates how indigenous inventory from one culture can benefit understanding of personality in other cultures and the search for universal and culture-specific personality traits.

Paper 1

"The external correlates of CPAI-2 traits in Asian and American college student samples: The case of openness and interpersonal relatedness"

Cheung, Shu Fai (University of Macau, Macau, China)*
Cheung, Fanny (The Chinese University of Hong Kong, Hong Kong SAR, China)

Two inventories to measure personality were administered: the CPAI-2 Form B personality scales, and the NEO-FFI for measuring personality factors from the Five-Factor Model, were administered to samples of college students from Mainland China (N=349), Taiwan (N=334), Hong Kong (N=364), South Korea (N=472), Japan (N=300), and America (N=305). Several self-report external correlates were also administered to measure aspects related to openness and interpersonal relatedness. Variety of friends, number of places visited, and variety in music preferences are expected to be related to openness. Another group of indicators are frequency having quarrels with family members or friends, giving gifts to family members or friends, and attending social functions such as weddings and funerals, which are expected to be related to interpersonal relatedness traits such as renqing (relationship orientation) and harmony. The replicability of the CPAI-2 factor structures have been examined in previous presentations. In the present presentation, the utility of personality scales in predicting these external behavioral correlates across samples from diverse cultural backgrounds will be examined. Specifically, the similarities and differences of the various predictors across samples and the additive utility of the CPAI-2 scales over and above Big Five will be presented. The theoretical implications of cross-cultural similarities and differences in the selected cultures will be discussed.

Paper 2

"The Cross-Cultural Personality Assessment Inventory-2: Structure, reliability and validity in an indigenous Dutch and a Chinese-Dutch student sample in The Netherlands"

Born, Marise (Erasmus Unviersity, Rotterdam, Netherlands)*

This presentation reports on an study into the generalizability of the personality structure underlying the CPAI-2 to a sample of indigenous Dutch (N=200) and Chinese-Dutch (N=150) university students. One of the CPAI-2 factors is Interpersonal Relatedness, which specifically attends to the interpersonal qualities of a person. This factor is different from any of the Big Five factors. Until now, support has been found for the Interpersonal Relatedness factor as a separate factor in Chinese, Asian American, Korean and Japanese people. Whether this factor and its subfacets are meaningful in Western, more individualistic samples has yet to be demonstrated (Heine & Buchtel, 2009). The present study examined this issue for The Netherlands, as an exemplar of an individualistic Western culture. To this end, the CPAI-2 translation was executed by integrating two separate Dutch translations by Chinese-Dutch bilinguals from the Chinese original and a translation from the English version into Dutch. A professional native Chinese speaking translator back translated the resulting Dutch translation into Chinese. Validation of the inventory was done by correlating it with the Hexaco-PI-R, which next to the Big Five includes a sixth factor, namely Honesty-Humility (De Vries et al., 2008). Its added value above the Hexaco-PI-R in predicting several meaningful student-related criteria was looked into. All analyses were done for both samples separately. Findings are discussed within the context of growing multiculturalism in Europe and the possible significance of Interpersonal Relatedness factor, originating from a non-Western culture.

Paper 3

"Personality traits, vocational interests, and career exploration: A cross-cultural comparison between Hong Kong and American university students"

Leong, Frederick (Michigan State University, USA)*
Fan, Weiqiao (Shanghai Normal University, China)
Cheung, Shu Fai (University of Macau, Macau, China)
Cheung, Fanny (The Chinese University of Hong Kong, Hong Kong SAR, China)

This study compared the pattern of relationships among personality traits, vocational interests, and career exploration between 392 Hong Kong and 369 American university students. The first hypothesis predicted differential contributions of the universal and indigenous personality dimensions based on the Cross-cultural (Chinese) Personality Assessment Inventory (CPAI-2) to career exploration of the Hong Kong and American students. A second hypothesis predicted that vocational interests of participants mediated the association between their personality traits and career exploration. Significant cultural differences were found between the personality predictors for Hong Kong and American students, supporting the first hypothesis. In addition, SEM results supported two significant mediator sub-models in the Hong Kong and American samples, respectively. The implications of personality traits and vocational interests to university students' career exploration were discussed.

Paper 4

"The CPAI-2 as a predictor of leadership behaviors and outcomes in Romania"

Iliescu, Dragos (SNSPA-FCRP Bucharest, Romania)*
Ion, Andrei (Testcentral Bucharest, Romania)

The Cross-Cultural (Chinese) Personality Assessment Inventory-2 (CPAI-2) was administered in Romanian on a sample of N=428 participants: 189 males (44.2%), ages from 18 to 57 years (m=30.6, SD=9.1). Together with the CPAI-2, the participants also completed two measures of leadership behaviors: the Leadership Opinion Questionnaire (LOQ, Fleishman, 1989) and the Multifactor Leadership Questionnaire (MLQ-5x, Avolio & Bass, 2004). Every participant was also assessed by a number between 1 and 5 observers with the help of the observer-evaluation forms of the two leadership measures: the Supervisory Behavior Description Questionnaire (SBD, Fleishman, 1989) and the MLQ for followers. Institutional "objective" scores of leadership performance were further collected based on the performance appraisal system of the respective organizations. These performance data, as well as the leadership outcome scales of the MLQ were treated as criteria in an hierarchical regression model which found that the CPAI-2 has a positive predictive power as a measure of leadership. Further factor-analytic approaches were employed to structure the collected data. Results show that the CPAI-2 scales contributes with incremental validity to the model, over and above the dimensions of transformational leadership as measured by the MLQ and the dimensions of focus on objectives vs. interpersonal interaction, as measured by the LOQ/SBD.

Paper 5

"Vietnamese version of the CPAI-Adolescent"

Dang, Minh (Vietnam National University, Vietnam)*
Weiss, Bahr (Vanderbilt University, USA)
Tran, Nam (Vanderbilt University, USA)
To, Hanh Thi (Vietnam National University, Vietnam)

This study has three phases. In Phase 1, the content of the CPAI was translated into Vietnamese. As part of this process, both the Chinese and English versions of the CPAI were translated separately into Vietnamese, and the two versions compared. In Phase 2 of the study, the factors of the CPAI will be analyzed in separate focus groups of (a) psychologists and other mental health related professionals, (b) parents of adolescents, and (c) adolescents. These focus groups will conceptually analyze the factors of the CPAI to determine if there are additional factors or constructs in Vietnamese culture that need to be included to comprehensively assess Vietnamese adolescents' personality. In Phase 3 of the study, the CPAI-VN will be administered to large sample of Vietnamese adolescents. Results first will be analyzed using confirmatory factor analysis (CFA) to determine if the hypothesized factor structure provides adequate fit for the empirical structure of the data. If the empirical structure differs significantly from the hypothesized structure, modification indices within the CFA analysis will be used to improve the fit of the structure. If the initial fit is adequate, then the data collected in Phase 3 be used to develop preliminary norms; if the initial fit is not adequate, an additional sample will be collected to cross-validate the modified factor structure, and these data will provide preliminary norms.

S23 08:30-10:00 Room201

Reconsidering Messick's facets of validity and their implications for demonstrating best practice in assessment

Chair

Burke, Eugene (SHL Group Ltd., UK)

Symposium Abstract

This symposium brings together a number of leading researchers and practitioners across a diverse range of verticals in the testing industry from broader HR practice to test reviewers and trade associations, to developers and providers of tests in the educational and employment sectors as well as providers of services in delivering secure testing programmes. The focus of the symposium is the seminal work of Samuel Messick (see below) in which he developed a taxonomy of six aspects of validity to be used in evaluating the quality of tests and other assessments: content, substantive, structural, generalisability, external and consequential. Each of the symposium members has been asked to choose at least one of these aspects of validity and to discuss their chosen aspect in the context of practical examples of best practice, where possible highlighting the best practice issues for global testing and assessment programmes as well as emerging programmes in China.

Paper 1

"Content"

Harris, William G. (Association of Test Publishers, USA)*

The work of Messick has continued to guide the measurement community's understanding of test validity. Messick championed successfully the conceptual argument that the tripartite division of validity into criterion-related, content and construct was inefficient and confusing. By the late 1980s it was generally accepted that "all measurement should be construct-referenced." Although this perspective has gained currency, the confusion around the content-construct relationship has remained. This presentation examines content as an integral thread of the test construction process and not as a discrete validation event. The value of a test relies heavily on the content relevance and content coverage, but neither of these domain characteristics convert directly to a test score or provide sufficient information upon which to defend test score inferences. Building a valid test is a multistage procedure that includes important stages such as psychological theory, prior research and operationally defined domain. These and other procedures are systematically introduced into the development process and form the foundation for creating content and eventually a valid test. The presentation will highlight the areas of content-construct confusion and apply Messick's conceptual framework to resolve such confusion.

Paper 2

"Substantive and structural"

Burke, Eugene (SHL Group Ltd., UK)*

This paper will address two concerns: whether an assessment has a clear a priori theoretical and measurement structure; the types of evidence that support the theory of the test as well as how well that theory operates under different conditions such as language and culture, as well as different candidate populations and testing contexts (e.g. low takes versus high stakes). Both maximum performance measures (cognitive ability tests) and typical performance measures (principally personality scales) are addressed. Issues addressed related to maximum performance measures are: • The relative influence of power and speed conditions in a test (and some fallacies related to speed versus accuracy in cognitive ability tests); • The contribution of item facets or radicals to the performance of test items; • The performance of test items in different languages and geographical populations (including Chinese as well as European samples). Issues addressed related to typical performance measures are: • Equivalence of item and scale properties under low stakes conditions such as trials in development versus high stakes conditions such as employee selection programmes (i.e. the faking problem); • Equivalence of items under different cultural conditions, whether they matter and what can be done to address them.

Paper 3

"Generalisability"

Rudner, Lawrence (Graduate Management Admissions Council, USA)*

This presentation focuses on cross-cultural considerations in evaluating validity generalization. A critical question for any test user is whether the existing validity evidence can generalize to a new situation without further study. Two aspects of this validity generalization question will be discussed: • How can a publisher document that the content of a test continues to adequately sample the intended domain?; • How can a publisher document that it is reasonable to expect the results of multiple past correlational studies to apply across cultural settings? Approaches to these questions employed with the GMAT® will be presented. More than 50% of GMAT test takers are citizens of countries other than the United States. The discussion on the content question will primarily address the need to examine the test blueprint. Is the mix of questions appropriate across cultures? The discussion of past correlational studies will primarily address the use of meta-analysis to determine whether the past studies are situation specific or whether the results do indeed generalize. Methods of aggregating studies and data from subgroups across studies will be presented.

Paper 4

"External"

Sun, James Jian-Min (School of Labor and Human Resource and Department of Psychology Renmin University of China, China)*

This presentation provides a perspective with evidence on cross-cultural issues in evaluating the external validity of psychological test. The validity issue in a cross-cultural setting becomes more complicated. Three main factors come into consideration when we talk about external validity cross-culturally: • What do we mean by external?; • What criteria should we establish for external validity study in different cultures and how could we verify the external validity of a single test developed in a specific culture?; • How could we define the differences among cultures which are meaningful for test validity studies? Possible solutions and approaches will be presented with examples of attitudinal test developed in western cultures but are becoming popular in China in recent years, such as transformational leadership, big five personality, core self-evaluations. Evidence with validation studies of the above tests in Chinese context will be presented with a discussion from John Berry's etic vs. emic perspective in cross-cultural research. Utility analysis will also be used as an evidence of external validity.

Paper 5

"Consequential"

Geisinger, Kurt. F. (The University of Nebraska-Lincoln, USA)*

There are a number of traditional ways that those consequential aspects of test use that impact validity have been identified. In the USA, the primary incidence is in predictive, selection, and/or hiring decisions when racial and ethnic group differences in test scores have impacted ultimate decisions and, in many cases, have led to adverse impact upon underrepresented groups. In educational testing, such impacts have also traditionally been found in criterion-referenced tests such as high school graduation tests. More recently, however, under No Child Left Behind, schools have been adversely impacted when their annual test results fail to meet expected standards. Schools with higher proportions of underrepresented groups are in many cases more vulnerable to these findings. Similarly, test takers who are disabled are often found to be underrepresented when test scores are used to make decisions. These factors will be discussed from the vantage of test validity, the extent to which unexpected or unplanned consequences should be a factor in making test use decisions, and guidelines that should guide proper test use in such situations.

Paper 6

"Messick's other facet"

Foster, David (Kryterion Inc., USA)*

In addition to helping measurement professionals understand the sources of evidence in construct validity Samuel Messick introduced and described the important concept and effects of construct-irrelevant variance (CIV) as a major threat to that validity. CIV makes a test easier or more difficult than it should be, irrelevantly raising or lowering test scores, and always creating adverse consequences for test takers, testing programs and other stakeholders. Some of these sources of CIV have existed since the beginning of modern testing, while others have been introduced more recently. This paper will describe some of these sources of CIV, will present research on their effects, and will provide guidance on what can be done to remove them.

S24 10:00-11:30 LT3

The development and validation of foundation measures of intelligence, creativity, personality, and temperament in Brazil

Chair

Wechsler, Solange (Psychology Institute, Pontifical Catholic University at Campinas, Brazil)
Oakland, Thomas (University of Florida, USA)

Symposium Abstract

Prior to the mid 1990s, psychological tests in Brazil remained almost unchanged during a 30-years period. Tests typically were acquired from other counties, mainly from the U.S., and translated into Portuguese, often with no studies related to their adaptation, validation, and norming in Brazil. As a result, the tests were criticized and discredited for being irrelevant for use in Brazil, including for the diagnosis of psychological disorders. This situation changed drastically through the efforts of Brazilian scholars from various universities throughout the country. A number of scholars created laboratories for test development at their local universities. Furthermore, they saw the need to create a national associated dedication to the promotion of sound methods for developing, evaluating, and using tests. The purpose of this symposium is to highlight efforts in Brazil that lead to the development of measures of intelligence, creativity, and personality.

Paper 1

"Advances in intelligence and creativity assessment in Brazil"

Wechsler, Solange (Psychology Institute, Pontifical Catholic University at Campinas, Brazil)*

Psychological tests in Brazil have undergone an incredible development the last six years. This was due to the fact that a federal regulation was passed in 2003 requiring that all tests being used in the country had to present empirical data informing their validity and reliability to the country, as well as current norms for Brazilians. All tests being used were impacted by these measures, leading to the creation of new measures as well as reviewing and adapting tests originated from other nations. This decision toward the improvement of test quality is not found in any other South American country. An overview of the tests that were approved, for use under these criteria will be presented. Emphasis will be given to intelligence and creativity measures, demonstrating their growth resulting from more emphasis on research on construction and adapting measures that are relevant to the country. Tests for children up to adults will be considered, emphasizing the need to adapt instruments created in another culture instead of just translating them. Language issues will be stressed as well as the difficulties involving in assessing verbal content in countries where there are no standard tests on educational achievement. In the other hand, socioeconomic status, which impact children from these nations, will be considered as a main difficulty presented when comparing their performance on tests with those of children from developed nations of the same age. Values issues are also to be considered, especially when assessing creativity. The expression of creativity tends to vary according to the values related to certain behaviors existing in each culture. Thus, certain types of creative behaviors will be more relevant to be identified in an environment than others. These issues bring implications for assessing gifted children as well as creative leadership among adults.

Paper 2

"Personality and positive psychology in Brazil"

Hutz, Claudio (Psychology Program, Federal University of Rio Grande do Sul, Brazil)*

The first paper about the Big-Five Factor Personality Model was published in Brazil in 1998. In the last decade we made solid progress and we have today two batteries to assess the model: a Brazilian version of the NEO-PI and a native scale that was developed from markers used in modern Brazilian Portuguese to describe personality traits. Both batteries present medium to high correlations. Before this native version of the Big-Five battery was developed, scales to assess each of the five factors we developed and validated. Thus we have now independent scales to assess Neuroticism, Extraversion, Agreeableness (in Portuguese it is called Socialization), Conscientiousness, and Openness. The availability of these instruments in Brazil allowed researchers to conduct studies in several fields. However, one field of growing interest in Brazil, Positive Psychology, also lacks instruments to measure most, if not all of its constructs. This is also a new area in our country. The first publication, a paper about resilience, was in 1996. Only very recently publications about subjective well-being appeared in Portuguese. We lack research relating personality characteristics, subjective well-being, hope, optimism, life events. Also, researchers want to examine the role of parental styles and educational practices in adolescents well being, self-esteem. The list of possible research projects is huge. Much of this research has already been conducted in the United States and in some other countries but it is necessary to verify if the cultural differences between Brazil and the US do not affect the development of these variables. We are now in the process of adapting instruments to measure Life Satisfaction for adults and for children, scales to assess positive and negative affects, scales for optimism, for hope, for self-esteem, and also a scale to assess life events in a somewhat different way. We are asking subjects to list events and to rate them, in a Likert scale, from not positive at all to very positive and from not negative at all to very negative. We found out that many events can be experience as positive and negative at the same time and this is bringing some interesting new perspectives to our findings. In conclusion, this presentation will provide information about valid scales available in Brazil to assess personality in the Big-5 model, valid scales to asses Life satisfaction, Positive and Negative Affect, Optimism, Hope, and Self-Esteem. It will also give a brief overlook of research relating Personality characteristics, Subjective Well being and life events

Paper 3

"The development of an adult measures of temperament"

Oakland, Thomas (University of Florida, USA)*

Interest in temperament may be as old as recorded history. Hippocrates was one of the first to speculate about temperament in 350 B.C. in his On the Nature of Man. The importance of temperament also is found in the writings of philosophers, including Plato, Aristotle, Galen, Bruno, Hume, Voltaire, Rousseau, Locke, and Kant. Jung advanced a more contemporary temperament theory, one that helped launched considerable research and test development. Temperament refers to stylistic and relatively stable traits that subsume intrinsic tendencies to act and react in somewhat predictable ways to people, events, and stimuli. Temperament traits generally are characterized as predispositions to display behaviors, a blueprint for them, with no assurance that people, events, and stimuli always will elicit the same temperament behaviors. Temperament traits appear early in life and thus are assumed to have a biological origin, one tempered both by one's environment as well as personal choice. Age and gender also are assumed to influence temperament. This paper discusses the development of an adult scale of temperament and compares Brazilian data with adult data from other countries.

Discussant

Byrne, Barbara (University of Ottawa, Canada)

S25 10:00-11:30 LT4

Chinese neuropsychological assessment

Chair

Chan, Agnes Sui-yin (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

Symposium Abstract

For the past ten some years, several ecologically validated Neuropsychological assessments have been developed by Chinese psychologists in either Hong Kong or China. For instance, the Hong Kong List Learning Test and Chinese version of Mattis Dementia Rating Scale have been published and used in Hong Kong for about ten years, and became two primary neuropsychological assessment tools for Chinese population. In addition, some tests on frontal lobe function were also developed in the past some years. A child neuropsychological assessment battery has been recently developed by a special work task under the Hong Kong Neuropsychological Association. The present workshop will focus on the recent development of neuropsychological assessment for Chinese population, and discuss the future development in this area.

Paper 1

"The Chinese version of Mattis Dementia Rating Scale in differentiating Dementia of Alzheimer's type in Chinese population"

Cheung, Mei-chun (Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong SAR, China)*

Chan, Sui-yin, Agnes (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

The talk aims to examine the clinical validity and applicability of the Chinese version of Mattis Dementia Rating Scale (DRS) for the Chinese elderly in Hong Kong. The scale is found to have good reliability with internal consistency ranging from 0.7 to 0.9. Its significant correlation with the Chinese version of Mini-Mental State Examination (CMMSE) suggests satisfactory construct validity of the scale. The discriminant validity of the Chinese version of DRS in differentiating Alzheimer's Disease (AD) patients and Normal Control (NC) elderly is also supported by the Receiver Operating Characteristics analysis. Several cutoff points will be presented for different clinical or research applications, and formulas to adjust for the age and educational level of the DRS total score will be provided. The Initiation / Perseveration and Memory subscales are suggested to be an abbreviated version of the Chinese version of DRS for a quick screening with adequate classification rate (91.9% overall correct classification rate). The present results also show the association between the impairment of the Chinese version of DRS performance and the progress of the Alzheimer's Disease.

Paper 2

"Clinical values and applicability of the Hong Kong List Learning Test for Chinese population"

Sze, Sophia, Lai-man (Integrative Neuropsychological Rehabilitation Center, The Chinese University of Hong Kong, Hong Kong SAR, China)*

Chan, Agnes S. (Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China)

Cheung, Mei-chun (Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong SAR, China)

The Hong Kong List Learning Test (HKLLT) is a standardized verbal list learning test with norm aged from 6 to 95 years old, which was developed and has been repeatedly validated in past decade among Chinese population. It is composed of two lists of 16 two-character Chinese words, with three learning trials, two delay recall trials and a recognition trial. While the words in one list are randomly arranged (random condition), those in the other are semantically clustered (blocked condition). It has been found to be a sensitive measure in differentiating memory deficits associated with various etiologies (e.g., Alzheimer's disease, Schizophrenia, Depression, Autism) and pathologies (temporal lobe versus frontal lobe damage) of brain disorders. The deficits in memory recall and strategy measured in the test are also found to vary with the severity / progression of some brain disorders, of which the profiles are sensibly associated with the possible pathological involvement and changes of the disorder. The presentation will compare the memory profiles of patients with Alzheimer's type of dementia, Schizophrenia and/or Autism as illustrative examples to demonstrate the utility of HKLLT for differential diagnosis. In addition, the clinical significance of HKLLT as a mean to provide insights into the direction of potential memory training will also be briefly presented.

Paper 3

"A normative study of neuropsychological profile in children and adolescents in Hong Kong: Preliminary findings on children normative profile"

So, Cheryl (Kwai Chung Hospital, Hong Kong SAR, China)

Chang, Sonia (Kwai Chung Hospital, Hong Kong SAR, China)*

In Hong Kong, normative data on neuropsychological profile of youngsters in normal population and in those with various inborn and acquired disorders are lacking. Most local studies on the neuropsychological assessment in young population have not recruited school-aged children (i.e., 6-12 years old). Therefore, there is a dearth of data on neuropsychological functions in children in Hong Kong. The purpose of the study is to examine the cognitive abilities in children and adolescents, including attention control, memory, language, reasoning and organization skills. This presentation will focus on the preliminary findings on the normative profile of Cantonese-speaking school-aged children in Hong Kong. Cross-cultural differences will be highlighted.

Discussant

Chan, Agnes Sui-yin (Department of Psychology, The Chinese University of Hong Kong, China, China)*

Monday, 19th July 2010

C01 11:00-12:30 LT4

C01-1

Enhancing interpretation of an individual student test score profile: Parametric and non-parametric approaches to estimate profile reliability

Arce-Ferrer, Alvaro (Pearson, USA)*

Abstract

This paper proposes two approaches –one non-parametric and the other parametric– to estimate profile reliability for an individual. Profile reliability for an individual is formally understood as a ratio of two variances: one tapping consistent item response patterns and the other tapping variability of test scores in the individual profile. Profile dispersion, the ups and downs in an individual test score profile, is a cornerstone piece of information when interpreting test scores and it is used in the derivation of the two reliability indices. Both approaches rely on a measure of the number of discrepancies between observed and expected vectors of item responses, but differ on its estimation. Whereas the non-parametric approach follows Guttman rules to derive expected vectors, the parametric approach derives expected vectors with an item response theory model. The paper first presents derivations of the two approaches. The paper moves to summarize the performance of the two indices with synthetic data from a Monte Carlo simulation with three manipulated facets – profile shape, measurement dependency, and measurement precision. In a following section the paper summarizes performance of the two approaches with empirically derived profiles from psychological and educational measurement. The paper ends with a discussion of the findings and recommendations for practitioners and researchers.

C01-2

5-Point or 6-Point? It is a Question

Li, Xuhong (Fudan University, China)*
Liang, Xiaoya (Fudan University, China)

Abstract

Survey researchers have long been concerned with whether providing a neutral middle point response will cause satisficing or non-response (Krosnick, 1999). It is also believed people are more likely to endorse the "Golden Mean" and avoid the extremes in Confucius societies (Behling and Law, 2000). Central tendency may become an inevitable response error when we use odd-point Likert-type rating scales in survey research. To investigate to what extent central tendency error occurs and what possible procedural and statistical remedies we can use to minimize its impact, we designed a controlled field experiment to explore how the "Golden Mean" affects subjects' responses to an odd number Likert scale. We split our subjects, 700 R&D personnel, into halves, with half responding to odd-numbered (e.g. 5 or 7-point) scale, and the other half to an even-numbered (e.g. 4 or 6-points) scale, to find out whether there is any difference. The comparative experimental results and their implications for considering cultural characteristics in adopting Likert-type rating scales will be discussed.

C01-3

One statistical technique may not be enough: Comparing measurement of psychological attributes across countries

Shulruf, Boaz (University of Auckland, New Zealand)*
Zeng, Min (Hong Kong University, Hong Kong SAR, China)
Watkins, David (Hong Kong University, Hong Kong SAR, China)
Hong, Fu (Nanjing University, China)

Abstract

The aim of this study was to compare the impact of the statistical techniques on measurement outcomes of Collectivism and Individualism scales across two countries, New Zealand (NZ) and China, and five ethnic groups. Attributes of Collectivism and Individualism (AICS) were compared across undergraduate students from Mainland China and New Zealand. Two statistical techniques (t-test and k-mean cluster analysis) were used to compare the samples and identify differences. The mean scores on both Collectivism and Individualism for the Chinese sample were lower. These results suggest that mean scores (e.g., t-test) of psychological measures do not provide enough information for international comparison and are therefore more susceptible to response biases than population classification-based analysis (e.g., cluster analysis). Implications for other psychological measures are discussed.

C01-4

Which is better? 5-point or 7-point scale?

Wu, Joseph (Department of Applied Social Studies, City University of Hong Kong, Hong Kong SAR, China)*
Lo, T. Wing (Department of Applied Social Studies, City University of Hong Kong, Hong Kong SAR, China)
Liu, Elaine S. C. (Department of Applied Social Studies, City University of Hong Kong, Hong Kong SAR, China)

Abstract

There is a general view that offering more alternatives on a rating scale could provide a more accurate measure of opinions/attitudes. However, there is also another view that a lengthy response scale might impose an unnecessary cognitive burden on respondents. To maintain a balance between these contradicting views, 5-point and 7-point Likert scales are commonly adopted among the odd-point scales. In this paper, relative merits of these two response formats were compared using data of 150 university students responding to an indigenous self-esteem instrument. As expected, the 7-point scale yielded wider response latitude and more scattered distribution than the 5-point scale. However, scores on the 5-point scale were more reliable than those on the 7-point scale in terms of internal consistency. In examining effects of age and gender on self esteem, using either scale led to similar findings which suggested that validity might not be compromised in using a shorter response scale. Implications for choosing the number of alternatives on a rating scale were discussed.

C01-5

The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data

Nordstokke, David W. (University of Calgary, Canada)
Zumbo, Bruno D. (University of British Columbia, Canada)*
Cairns, Sharon Lynn (University of Calgary, Canada)
Saklofske, Donald H. (University of Calgary, Canada)

Abstract

Many assessment and evaluation studies use statistical hypothesis tests, such as the independent samples t-test or analysis of variance, to test the equality of two or more means for gender, age groups, cultures or language group comparisons. In addition, some, but far fewer, studies compare variability across these same groups or research conditions. Tests of the equality of variances can therefore be used on their own for this purpose but they are most often used alongside other methods to support assumptions made about variances. This is often done so that variances can be pooled across groups to yield an estimate of variance that is used in the standard error of the statistic in question. The purposes of this paper are: (a) to describe a new nonparametric Levene test for equal variances (Nordstokke & Zumbo, 2007; in press) that can be used with widely available statistical software such as SPSS or SAS, and (b) to investigate this test's operating characteristics, Type I error and statistical power, with real assessment and evaluation data. To date, the operating characteristics of the nonparametric Levene test have been studied with mathematical distributions in computer experiments and, although that information is valuable, this study will be an important next step in documenting both the level of nonnormality (skewness and kurtosis) of real assessment and evaluation data, and how this new statistical test operates in these conditions.

C02 14:00-15:30 LT4

C02-1

Validity research on situational judgment integrity tests for managers in the Chinese culture

Chen, Lijun (College of Public Administration, Zhejiang University, China)*

Abstract

According to previous research, integrity was selected as the first moral character that an entrepreneur must have. It is thus very important to develop a proper instrument to evaluate managers' integrity for personnel selection. But when we use the traditional psychodynamic measures (e.g., Covert and Overt paper-and-pencil tests) to evaluate an employee's integrity in the Chinese culture, there are remarkable social desirability effects. Our study focused on developing a situational judgment integrity test for the Chinese culture and testing its validity. By adopting the construct-oriented and job-oriented method, we developed the manager's Situational Judgment Integrity Test (SJIT). Then by using three different samples of 176 employees and undergraduates, we tested the construct validity and criterion-related validity of the SJIT in the Chinese culture. Exploratory Factor Analysis results demonstrated that the 3-dimension construct (Honesty & Accountability, Regulation Compliance & Clearness, Justice & Fairness) which was developed in the test, fit the data. Applying the Big-Five personality scale, we found that integrity scores were significantly related to the scores on the factor of conscientiousness. Thus, the results suggested that the SJIT has good construct validity. Higher predictive validity was also verified by other-rating criteria (regarding Criteria 1, r=0.50; and regarding Criteria 2, r= 0.31). No significant difference on SJIT test scores was found between the employee group that took the SJIT under the condition of personnel evaluation and the employee group in the T&E environment; a social desirability effect was not observed. Implications and some unresolved questions are discussed.

C02-2

You want me to do what? Changing the traditional role of subject-matter experts in test development

Eatchel, Nikki (Prometric)*
Ro, Shungwon (Prometric)
Miller, Bonnie (Prometric)

Abstract

Global testing organizations are faced with multiple challenges. One such challenge involves the participation of subject-matter experts (SMEs) in lengthy item review processes. This becomes particularly challenging for content categories involving advanced expertise requiring international SMEs who are in high demand and have little time to commit to such projects. Though methods for item writing have become more flexible, item review processes tend to follow traditional group reviews that require significant SME commitment. The level of commitment required can lead to scheduling problems (as experts cannot commit within required timeframes) and quality issues (as experts who can commit may not constitute the right mix of backgrounds). This paper describes a project in which a more flexible review process was implemented involving a series of independent reviews and collaborative feedback amongst SMEs. The goal was to increase the breadth and depth of the subject-matter experts involved, while evaluating the psychometric quality of the resulting items in comparison to those reviewed in a traditional group setting. A comparative analysis of both sets of items will be presented, including benefits and drawbacks of the different approaches. As global test development is now the standard, identifying psychometrically sound, alternative processes to traditional group settings is a necessity, and this paper's findings could have a significant impact on future test development processes.

C02-3

Three countries, one exam: How one client's business challenge resulted in a new approach to test development

Eatchel, Nikki (Prometric)*
Li, Dongyang (Prometric)
Hampton, O'Neal (Prometric)

Abstract

The test development industry is plagued by the push and pull of two consistent demands: (1) create psychometrically sound, high-quality exams, and (2) keep development costs reasonable for the organizations involved. When one international client identified the need to reduce costs by

developing examination content in two countries (neither of which was the source country for the exam), a team embarked on a ground-breaking design to ensure quality and validity for the candidate population taking the exam, while utilizing subject-matter experts with varying backgrounds in specified geographic locations. This paper outlines the lessons learned from the process, the key design elements that were critical to a valid end product, the obstacles that arose, the necessary psychometric evaluation employed, and the performance results of the end product – including a comparative analysis of items in the same content areas developed by SMEs in the source country alone versus the two alternative countries requested. The implications of the project results are substantial for the test development industry – particularly as global clientele seek to utilize varying SME populations. The information outlined in the presentation will help pave the way for a cost-effective, global approach to test development.

C02-4

16-year-old students' vocational interests in 65 countries using the PISA surveys: Cross-national and between gender differences.

Rocher, Thierry (Université Paris Ouest Nanterre La Défense, France)
Vrignaud, Pierre (Université Paris Ouest Nanterre La Défense, France)*

Abstract

Holland's vocational interest typology has been for half a century the prominent model for the assessment of vocational interests. Many studies about the cross-cultural validity of this model heave been published (Rounds & Tracey, Vrignaud, 1994). But these studies are in general conducted using specific samples, usually of high school or university students. The PISA surveys assessing 16 year-old pupils' literacy are conducted in more than 50 countries (OCDE and countries wishing to participate) on representative national samples. In the background questionnaire, there are several questions on the expectations of the pupils in terms of jobs. This information allows us to study vocational interests according to Holland's RIASEC typology. We analyzed these data using several statistical methods (ACP, hierarchical clustering). The results highlight the existence of differences in popularity of the six RIASEC types between countries. Moreover the study of the interaction between gender and countries show contrasting profiles of interest and popularity. These kinds of results on representative samples and so many countries has, to our knowledge, never been published before. We discuss the possibility of interpreting these differences in relation to the pupils' background and show their heuristic nature to test the universality of Holland's model.

C02-5

Investigating construct validity of the Group Climate Survey Short Form with multilevel-CFA, IRT and PLS

Wang, Y. Lawrence (National Changhua University of Education, Taiwan)*
Chen, Junesue (National Changhua University of Education, Taiwan)

Abstract

The purpose of the study is to investigate the multilevel (individual and group level) factorial structures of the Group Climate Survey-Short Form (GCQ-S) of group counseling by using three unique analytic models including multi-level confirmatory factor analysis, multilevel IRT and Partial Least Squares. Though the three factors of group climate including engagement, avoidance and conflict (MacKenzie, 1983) are significant predictors of success of procedural development and therapeutic effects of group counseling, studies have shown moderate to poor evidence of construct validity for the GCQ-S and have suggested that further investigations on the psychometric properties of the GCQ-S are needed. The authors investigated this validity issue by examining cross-sectional and longitudinal data to confirm the three-factor structure and to assess the invariance of the factors, respectively. An integrated sample from two sequential studies of group counseling, each consisting of nine sessions of group counseling was drawn, resulting in a total of 41 groups and 389 members, thus meeting the sample size requirement of multilevel analysis. Data were analyzed using MPlus for CFA and IRT (multilevel) and using Smart PLS for PLS (single level). The preliminary results from the multilevel CFA of cross-sectional data show that the sixth and the eighth items of the GCQ-S performed poorly, as they demonstrated non-significant factor loadings and explained little variance. Results from a cross-sectional study of IRT and PLS demonstrate somewhat similar results. The three-factor structure showed satisfactory fit indices when these two items were removed from the model. Results from analysis on longitudinal data will be discussed.

C02-6

Development of a test of undergraduate students' readiness for obtaining employment in China

Xu, Jian Ping (School of Psychology, Beijing Normal University, China)*
Yang, Min (School of Psychology, Beijing Normal University, China)
Wu, Lin (School of Psychology, Beijing Normal University, China)
Tan, Xiao Yue (School of Psychology, Beijing Normal University, China)

Abstract

The purpose of this study was to develop a test of undergraduate students' readiness for obtaining employment based on a competency model of excellent employees working in technical, financial, marketing, administrative, service and managerial jobs. We analyzed employment advertisements and documents using a qualitative method, to construct the test items. Five hundred seventy seven undergraduate students and 247 graduate students majoring in humanities and science and engineering courses respectively, from 11 different universities, were assessed by the test. The test was composed of 105 items and 21 subscales, which included communication skills, achievement orientation, learning capacity, client orientation, self-confidence, conscientiousness, initiative, professional dedication, ability to analyze and judge, ability to reason, integrity, innovation, relationship-establishing, influencing power, insight into interpersonal relationships, organizing and coordinating ability, teamwork spirit, encouraging capacity, executive ability, planning capacity and others. Reliability analyses, confirmatory factor analyses and correlation analyses showed that the reliability and validity of the scale achieved the psychometric criteria. Significant differences between the excellent-performing group and the average-performing group were found on total test scores and subscale scores including: learning capacity, self-confidence, initiative, planning capacity and ability to reason. The

reliability and validity of the test were thus strongly supported. The test could be effective to identify excellent students and could be an effective tool for undergraduate students for self assessment before employment and for career planning.

C03 14:00-15:30 LT2

C03-1

Tertiary education's challenge: Conducting fair and appropriate assessment in large foundation courses.

Clinton, Janet M. (University of Auckland, New Zealand)*
Exeter, Daniel (University of Auckland, New Zealand)
Leeson, Heidi (University of Auckland, New Zealand)

Abstract

Increasing economic constraints and globalised approaches to higher education mean many universities offer generic courses designed to target students from different faculties or universities, catering to a range of academic cohorts. This typically results in large classes with a diverse student population, hence increasing the difficulty in maintaining consistent achievement. This is the case in a New Zealand university with a large core population health course for students enrolled in Health Science, Nursing, Pharmacy and Biomedical Science programs all participating and taking part in assessments in the form of two MCQ tests during a year-long program. Statistically significant differences in achievement across the various cohorts of students in this course have been reported; however, it is not clear whether this is because of any bias in the items or prior achievement before entry. This study uses differential item functioning (DIF) analysis to investigate the items and dimensions of the MCQ exam, to better understand how 2,000 students from different programs respond on the assessments over two cohorts from two different years. DIF provides an indication of how the various cohorts respond to items and also gives valuable feedback for teachers and students. This paper demonstrates that DIF analysis is a useful means of understanding student responses to assessments in a large generic course. Finally, the implications and challenges for assessments are explored and suggestions for feedback to staff and students in such large generic courses are recommended.

C03-2

Setting the standard of English for entry level nurses

De Jong, John H. A. L. (Pearson Language Tests, United Kingdom/ VU University Amsterdam, Netherlands)
Li, Jinshu (Pearson Language Tests, United Kingdom)
Duvin, Helene (Pearson Language Tests, United Kingdom)
Zheng, Ying (Pearson Language Tests, United Kingdom)*

Abstract

Nurses whose first language is not English and who seek employment in the USA need to take a language test to demonstrate sufficient level of English. In cooperation with the National Council of State Boards of Nursing (NCSBN) Pearson conducted a standard-setting workshop on the Pearson Test of English Academic (PTE Academic). Standard-setting was conducted in three rounds: (1) establishing requirements, (2) evaluating difficulty of items and ability of candidates against the level descriptors of the Common European Framework of Reference for Languages (CEF) and (3) deciding on which items and which candidates represented the minimum standard. Members on a panel representative of the USA nurses community voted independently in the three rounds. Because the PTE Academic is scaled on the CEF, the difficulty and ability estimates from the panel expressed in terms of the CEF could be transformed to scores on the test reporting scale. High levels of agreement were reached among panel members and high correlations were found between candidate and item data from field tests and the evaluation of the panel. In a final round, panelists expressed their overall opinions on where the entry level nursing skills fell on the CEF scale. These judgements were again translated to the test reporting scale and were found to be in close agreement with the results from the item and candidate level third round decisions. The implication of this study is that the CEF provides a robust descriptive framework, allowing for test content-based decisions on minimal standards.

C03-3

An analysis of the high representation of scientific and mathematical disciplines in new zealand scholarship premier awards

Johnston, Michael (New Zealand Qualifications Authority, New Zealand)*

Abstract

The Premier New Zealand Scholarship, awarded annually to students performing at a high level in multiple secondary school subjects, has been historically dominated by students with successful results in mathematics and the sciences. Three possible explanations for this dominance are explored in the present paper. The first is that greater numbers of students undertake assessments in combinations of these subjects than in combinations of other subjects. The second is that there is a greater correlation in the cognitive demands of subjects within this group than is the case for other subjects. The third is that candidates undertaking mathematics and science are stronger scholarship candidates, on average, than students undertaking other subjects. The analyses show that all three explanations have some currency, although the correlation in cognitive demands is shown to be no greater within the mathematical disciplines and the sciences than it is within the humanities.

C03-4

Looking global - thinking local

Orchard, Sue (Comms Multilingual Ltd., England)*

Abstract

The topic of globalisation is gaining momentum in the world of associations and professional organisations. Up until now, such organisations have been content just to service their own membership within their own countries and they have done this very well. Why should associations now look at Global Certification? We now live in a world where there is unprecedented interconnection between different cultures. There is an increase in global alliances and supply chains and new markets and new products are emerging all the time. This interconnection requires people to have global skills which are transportable, and this is where professional certifications and tests that have a global acceptance are coming into their own. While

offering a certification or test in another language and/or country can create an expansion opportunity for an association and an opportunity to increase its global relevance, the journey to seizing such opportunities requires a sound strategy and the expertise to execute on it. Using actual case studies, this paper will discuss what needs to be considered when associations are thinking about offering their certification or tests in other countries. An outline of a process will be provided, which shows the steps that need to be taken in order to ensure successful acceptance and use of an adapted certification in the target market.

C03-5

A survey of test taker beliefs and experiences: A social psychology of assessment and testing

Pope, Greg (Questionmark)
Breithaupt, Krista (Medical Council of Canada, Canada)
Zumbo, Bruno D (University of British Columbia, Canada)*

Abstract

This session presents the results and observations of a comprehensive survey conducted in three countries in three contexts on examinees who have taken licensing, certification, and post-secondary exams. Examinees were asked a number of questions regarding their perceptions of the fairness of many assessment practices and issues common in the testing industry. Question topic areas included test security, reliability and validity of assessment scores, and test-taking medium. Analyses of results examining differences between and within licensing, certification, and post-secondary contexts will be presented and discussed. Demographic analyses regarding examinee perceptions of fair assessment practices will also be presented and discussed. The information presented in this session may be used by organizations to ascertain the impact of applying certain assessment techniques or approaches and how these techniques or approaches could impact examinee satisfaction, motivation, performance, and likelihood to appeal.

C03-6

Techniques for ensuring linquistic equivalence among forms of an academic self-concept scale rendered in eight languages

Michael, Joan J. (North Carolina State University, USA)*
Gerhold, Jennifer (Motorola, Inc., USA)

Abstract

The Dimensions of Self-Concept (DOSC) scale is a 96-item, self-reporting questionnaire comprising six factor subscales, each representing one of six hypothesized dimensions of academic self concept: Level of Aspiration, Anxiety, Academic Interest and Satisfaction, Leadership and Initiative, Identification vs. Alienation, and Stress (Michael & Smith, 1976; Michael, Smith, & Michael, 1989). Forms have been constructed in English for use with elementary and secondary school students, as well as with students in higher education, and with personnel in occupations requiring relatively high levels of formal education. Selected forms have been translated into Arabic, Chinese, Hebrew, Japanese, Korean, Portuguese, and Spanish using a translation and back-translation method developed by Michael et al. Translated forms of the DOSC have submitted to both construct and criterion-related validity studies, the results of which have been published in a number of professional journals. Evidence has indicated a relatively high degree of factorial invariance of scores across forms administered to individuals for whom their first language is that of the DOSC form. This presentation discusses the translation method used to ensure, as nearly as possible, the linguistic equivalence between versions of the DOSC.

C04 15:45-17:15 LT2

C04-1

Random effects models for meta-analytic structural equation modeling

Cheung, Shu Fai (University of Macau, Macau, China)*
Cheung, Mike (National University of Singapore, Singapore)

Abstract

Meta-analytic structural equation modeling (MASEM) is usually applied to synthesize research findings in path analysis and structural equation modeling. It can be used to validate the factor structure of tests or instruments across different settings. In the first step, correlation matrices are pooled together. Proposed models are then fitted on the pooled correlation matrix in the second stage. Currently, most methods in MASEM are based on the fixed-effects models. That is, the parameters are assumed to be the same in all studies. Limited studies have been conducted to investigate how random-effects models can be fitted in MASEM. The main objective of this study is to explore different types of random-effects models in the context of MASEM. Real examples will be used to illustrate how different random-effects models can be implemented. Advantages and disadvantages of different models will be discussed.

C04-2

Testing for measurement invariance in covariance-based and variance-based structural equation modeling: Comparisons of multi-sample analysis using smart-PLS and Mplus

Chiou, Hawjeng (Department of Business Administration, National Central University, Taiwan)*
Fu, Sian-Jhih (Department of Information Engineering, National Central University, Taiwan)
Lin, Pi-Fang (Department of Education, National Cheng-Chi University, Taiwan)

Abstract

The measurement of unobserved constructs as well as their equivalence across samples is the most challenging of tasks for social scientists who are interested in performing cross-cultural comparisons. The most popular way for estimating the constructs is confirmatory factor analysis, a sub-model of structural equation modeling (SEM). According to the literature, however, there are two approaches to performing an SEM, the covariance-based approach (CSEM) and the variance-based approach (VSEM). VSEM is the preferred alternative to CSEM due to its advantages such as being distribution free and requiring a smaller sample size, etc. However, the ability to detect the presence or absence of measurement invariance is still unknown for the application of VSEM. By using a Monte Carlo simulation, this paper compares the ability to detect the measurement invariance of VSEM and CSEM on different conditions. For VSEM, the test for partial least squares pooled significance test for multi-groups using SmartPLS was used for testing the invariance of factor loadings and factor scores. For CSEM, multiple group analysis in Mplus was used for testing for the invariance of factor loadings and intercepts of latent factors using χ2 difference tests for a set of nested models. The current results revealed that VSEM is less successful at detecting the true differences across two samples; CSEM required relatively bigger sample sizes and normally-distributed indicators to achieve the optimal model fit. An empirical analysis using real data was processed to reproduce the simulation results.

C04-3

Using the multitrait-multimethod matrix approach (MTMM) for psychometric analysis tests with a two-entry table structure

Mitina, Olga (Moscow State University Lomonosov, Moscow City University of Psychology and Education, Russia)*

Abstract

The MTMM was developed by Campbell and Fiske (1959) to examine Construct Validity. Usually with MTMM, data are collected using the same test to measure a set of traits with different methods (e.g., different respondents from different points of view assess the same person) or by measuring the same traits with similar tests. But there are many tests which measure constructs which differ from each other by field of expression, types of activity or level of functioning. Taking this into consideration it is possible to organize all scales included in such types of test into a two-entry table. As a result, all scales from the same row and the same column have something in common from a psychological point of view. MTMM is very useful to test and interpret hypotheses in the process of creation and (or) adaptation of tests. We can study the constructs as results of interactions between aspects constituting rows and columns and draw conclusions about their construct validity. SEM is used to analyze data. In this paper this method will be illustrated using examples from the psychology of motivation (the motive implementation test by Kuhl (1997)) and social psychology (a newly-developed test measuring social tolerance Babaeva, Sabadosh, Mitina (2010)). In the first case according to the theory three motives – power, affiliation and achievement – are viewed against different levels of personality: behavior, emotion, attitudes, and intentions. In the second case, different types of tolerant (intolerant) behavior and attitudes toward different minority and marginal groups were measured. Data were collected in Russia.

C04-4

Ridge structural equation modeling with correlation matrices for ordinal and continuous data

Yuan, Ke-Hai (University of Notre Dame, USA)*
Wu, Ruilin (Beihang University, China)
Bentler, Peter (University of California at Los Angeles, USA)

Abstract

The paper develops a ridge procedure for structural equation modeling (SEM) with ordinal and continuous data by modeling polychoric/polyserial/product-moment correlation matrix R. Rather than directly fitting R, the procedure fits a structural model to R_a=R+aI by minimizing the normal-distribution-based discrepancy function, where a>0. Statistical properties of the parameter estimates are obtained. Four statistics for overall model evaluation are proposed. Empirical results indicate that the ridge procedure for SEM with ordinal data has better convergence rate, smaller bias, smaller mean square error and better overall model evaluation than the widely used maximum likelihood procedure.

C04-5

Testing the invariance of latent traits in multiple group analysis

Zhang, Zhiyong (Department of Psychology, University of Notre Dame, USA)*

Abstract

Evaluating measurement invariance (or differential item functioning, DIF) is critical for international testing. Traditional methods require the same set of test items to be used in all populations. After obtaining invariance, the latent traits can then be compared. This study proposes to test the invariance at the latent trait level so that the test items can be tailored according to each population. For example, authoritarian behavior might be the target latent trait in a study, but in one population authoritarian behavior might be indexed by measures of physical coercion while in another population it might be indexed only via verbal behavior measures. Thus, the latent traits are the same although the test items are different. We propose to use marker variables and the likelihood ratio statistic to test the invariance of latent traits across different populations. Through well-designed simulation experiments, we demonstrate the feasibility of testing latent trait invariance when tailored items are used. The study has important implications in international testing. For example, the design of studies can be innovative in building measurement batteries that are better matched to one's experimental aims. For example, in studies involving subgroup comparisons based on age, ethnicity, gender, etc., test batteries can be tailored to the subgroup while still measuring the same latent traits.

C05 15:45-17:15 LT3

C05-1

Computerized assessment of basic aptitudes in international pilot selection at the German Aerospace Center – Comparison of data from different countries

Albers, Frank (German Aerospace Center (DLR), Department of Aviation and Space Psychology, Germany)*

Abstract

The German Aerospace Center (DLR) has been conducting pilot selection for civil commercial airlines since the 1950s. For more than ten years now, required basic aptitudes and knowledge according to international regulations (JAR-FCL) have been assessed by means of computerized tests. The assessment is done either in test centers with fixed computer hardware or with mobile testing equipment. The ability domains, which are requirements for pilot applicants, and the tests used for assessment are

briefly introduced. These domains cover aspects like selective attention, spatial comprehension, memory function, psychomotor coordination, multiple task capacity and knowledge in different areas, e.g., English language. The selection strategy, which is similar to that of various international airlines, is sketched. Test takers' performance data from samples of three international airlines from different countries in Europe and Western Asia are compared. Differences and similarities between test takers from different nationalities within the samples will be compared on a number of criteria (performance levels, distributions and factor structures). Implications and future directions will be discussed.

C05-2

Studying interface effects on computer based testing: Does change in computer interface matter?

Arce-Ferrer, Alvaro (Pearson, USA)*

Abstract

Multinational assessment programs offering both computer and paper versions of a test face challenges to document score comparability between test administration media, and to assess score comparability between computer interfaces. In this paper we focus on the latter challenge. This paper experimentally studied the presence of interface effects and their effects on students and item performance by developing two competing interfaces for a common assessment in which items are presented one at a time. The interfaces differ by design decisions on interactivity with the computerized test (i.e., navigation, skipping items and marking answered items for review). Two experimental data sets were collected under the random groups and the anchor test designs to increase generalizability of results. The former design involves random assignment of students to either of the two interfaces, and the latter design involves groups of students taking the paper version of the assessment (i.e., an anchor test) in addition to administrations of either of the two interface conditions. The experimental study subsumes a population of examinees, two test administration interfaces, two test modes of a target assessment, and two data collection designs. The analyses of the experimental data are performed for students and items with inferential statistics and item response theory approaches. The final paper summarizes effect sizes of interface effects on test and item level measures.

C05-3

Computer adaptive testing with multidimensional forced choice items to support selection and classification decisions

Chernyshenko, Oleksandr (Nanayng Technological University, Singapore)*
Stark, Stephen (University of South Florida, USA)
Drasgow, Fritz (University of Illinois at Urbana-Champaign, USA)

Abstract

Many organizations desire to use personality tools to support their selection and classification decisions. Such tools need not only be rooted in empirical research, but also be flexible in terms of test delivery options as well as resistant to socially desirable responding. In this paper, we describe our on-going effort to employ modern psychometric methods and computing technology to design Tailored Adaptive Personality Assessment System (TAPAS) to be used in large volume, high stakes testing contexts. TAPAS uses multidimensional pairwise preference items intended to discourage socially desirable responding. It can be administered in a computer adaptive format, which makes it possible to tailor item selection to each examinee, to increase test efficiency without sacrificing accuracy, and to control item exposure. We will describe the results of several simulation studies that compared score estimation accuracy for nonadaptive and adaptive tests involving different numbers of dimensions and test lengths. We will show that test length can be reduced by about 50% in operational settings using adaptive item selection. We will also present selected empirical results from studies involving TAPAS applications to assist with military selection and classification decisions.

C05-4

Computerized assessment of basic aptitudes in international pilot selection at the german aerospace center – implementation of computer based training tools and effects on test performance

Huelmann, Gerrit (German Aerospace Center (DLR), Department of Aviation and Space Psychology, Germany)*

Abstract

Computerized testing in personnel selection at the German Aerospace Center (DLR) was introduced in the late 1990s. At the same time the free and easy availability of information through the internet increased significantly. As a side effect of this development, specific information on tests could be spread easily among test takers, thus the issue of test security had to be re-evaluated. With an uncontrolled amount of more or less detailed information on tests used in pilot selection being available to an equally uncontrolled part of the tested population, the quality of the aspired diagnostic decision as well as testing fairness was jeopardized. As a response to this situation, computer based training tools (CBT) were developed and distributed freely to all test takers prior to their assessment. To really gain an increase in fairness, these training tools had to resemble the actual tests as accurately as possible.

- The history of computerized testing at the German Aerospace Center will be outlined.
- The evolution of computer based training tools and different training concepts to optimize test fairness will be illustrated.
- Large samples of test takers' performance data under different training conditions will be compared.

C05-5

A comparison between fixed and flexible content balancing methods in computerized adaptive testing

Leung, Chi Keung Eddie (Hong Kong Institute of Education, Hong Kong SAR, China)*

Abstract

Among the practical problems related to applications of computerized adaptive testing (CAT) in real-life testing programs, the problems of item exposure control and content balancing are most urgent. Overexposure of items can cause security problems while too much variation in content among adaptive tests may raise concerns about content validity. Content requirements can range from simple to complex. For simple content requirements, where less content is of interest, three easy-to-use techniques, namely, the constrained CAT (CCAT), the modified constrained CAT (MCCAT) and the modified multinomial model (MMM) are available. These three methods originally tackle situations where the number of administered items from each content area is fixed. In some testing programmes, the content requirements may not be so rigid. Instead of a fixed number of items, a flexible range of number of items from respective content areas would be allowed. To meet the needs of these testing programmes, we would introduce in this study three flexible content balancing methods by modifying the three existing fixed content balancing methods. The comparison between the performance of the flexible methods and the fixed methods was done through simulation studies. The results indicate that only the flexible form of the MCCAT has yielded significant improvements in both measurement accuracy and item usage while the flexible forms of the other two content balancing methods have not produced much gain. This paper will report in more detail about the development of the flexible content balancing methods and the performance comparison.

C05-6

Computerized assessment of basic aptitudes in international pilot selection at the German Aerospace Center - The validity of knowledge tests

Zierke, Oliver (German Aerospace Center (DLR), Department of Aviation and Space Psychology, Germany)*

Abstract

A validity study was conducted for two purposes: to evaluate the importance of knowledge tests for pilot selection and to validate the computerized knowledge tests from the German Aerospace Center. These tests were item bank based tests and covered mathematical, technical, and physical abilities and English knowledge. The temporal relation of predictor and criterion for knowledge tests will be discussed. Our criterion was a theoretical test of the Flight Training School where pilot applicants were admitted after passing the qualification. Regression models were calculated to measure the incremental validity of knowledge tests beyond cognitive tests. A further question addressed the comparison of the predictive validity from knowledge tests and school grades. Knowledge tests contributed 12% of incremental validity and were similar in value to school grades (17%).

C06 15:45-17:15 Room201

C06-1

Can assessment help enhance reading literacy? - Lessons from pirls

Choi, Hye-Jeong (University of Georgia, USA)*
Cohen, Allan S. (University of Georgia, USA)

Abstract

The Progress in International Reading Literacy Study (PIRLS) is a large-scale comparative study to monitor international trends in primary school reading achievement, in the fourth grade, on a 5-year cycle beginning in 2001. For this assessment program reading literacy is defined as “the ability to understand and use those written language forms required by society and/or valued by the individual.” Owing to the longitudinal design, it allows for comparing the performance of students across countries and across years within one country. In general, results were used to summarize proficiency for countries. To date, however, there has been no attempt to provide instructive information as to how to enhance reading literacy, which can be one of the most essential goals of assessment. The purpose of the current study is to propose an alternative model in response to the need for an informative assessment model. The model combines a Rasch model with a cognitive diagnostic model. This model is promising because it will provide information as to not only how much students perform, but also why some students have difficulty and how we can help students overcome those difficulties in reading in terms of atomic knowledge which constitutes reading literacy.

C06-2

The who and the why - Repeat testing patterns around the world on the GMAT exam

Defibaugh, Courtney (Graduate Management Admission Council, USA)*
Taliaferro, Hillary (Graduate Management Admission Council, USA)
Rudner, Lawrence (Graduate Management Admission Council, USA)
Talento-Miller, Eileen (Graduate Management Admission Council, USA)

Abstract

For some, the idea of sitting for a standardized test once is terrifying enough; doing it again would be considered only under the direst of circumstances. The GMAT exam is a computerized adaptive test that is administered nearly continuously around the world, with policies that provide for retaking the test within certain guidelines. Logically, people who think they will score higher on a subsequent trial, such as those whose scores were reduced for not finishing a section, may invest the time and money to try again. Previous research suggests that non-native English speakers tend to repeat more often. Challenges with the language and timing of the test may influence the decision to retake the test. This research explores whether cultural differences exist in repeat testing behavior. Specifically, differences among citizenship and language groups will be examined relative to gains observed, as well as number of repeats, reporting and cancelling of scores, and time spent on sections. Understanding differences among the motivations and experiences of examinees helps to ensure policies and availability of the test are appropriate for all.

C06-4

The use of pictorial supports as an accommodation for increasing access to test items for students with limited proficiency in the language of testing

Solano-Flores, Guillermo (University of Colorado at Boulder, USA)*

Abstract

This paper will report on an NSF-funded project that examines vignette illustrations (VIs) as a form of testing accommodation for English language learners (ELLs)—students who are developing English as a second language yet who are tested in English. VIs are pictorial supports intended to illustrate certain words or expressions from the text of items that may be uncommon or unfamiliar to ELLs. Based on semiotics, socio-cultural theory, and cognitive science, our conceptual framework defines the intended functions of images and a grammar of visual constituents. We have developed a procedure for systematically designing VIs for test items and for examining: illustrability (when should an item be illustrated?), criticality (what text elements should be illustrated?), complexity (what is the optimal set of VI constituents?), and universality (what visual elements may not be interpreted by all students in the same ways?). To evaluate the effectiveness of VIs, we use cognitive interviews to examine whether VIs help ELLs to better understand science items without altering the construct measured and whether ELLs and non-ELLs interpret vignette-illustrated items differently. Also, we use generalizability theory to examine the psychometric properties of vignette-illustrated items and, more specifically, the amount of score variation due to the main and interaction effect of student and two facets, item and presence-absence of VIs. The paper will discuss the potential of VIs as a valid, cost-effective, easy-to-implement testing accommodation in multilingual and multicultural contexts in which student language proficiency in the language of testing is a potential threat to test validity.

C06-5

Examining test speededness by native language

Talento-Miller, Eileen (Graduate Management Admission Council, USA)*
Guo, Fanmin (Graduate Management Admission Council, USA)
Han, Kyung (Chris) T. (Graduate Management Admission Council, USA)

Abstract

One consideration in the administration of a test used globally is the myriad backgrounds represented by the examinee population. It is appropriate to administer a test in English when the purpose of the test requires facility in that language, but beyond the content itself, native language may have an effect on the speed of responses to items. For a power test, with no time limits, there should be no effect of speed, but when time limits are necessitated by practical constraints, the effect of speed by native language should be examined. Combining all non-English languages may mask effects among different languages based on their degree of differences from grammatical conventions and the alphabet used in English. The availability of a large global audience allows examination of the effect of speededness for different language groups on the GMAT exam. The standardization method, which was previously adapted to assess differential speededness while accounting for differences in ability, will be used with the language groups. The technology used for the test allows examination of the effects of rapid guessing that may be necessitated by the time limits in addition to the items-not-reached criteria more typical of previous studies of differential speededness.

C06-6

Establishing test validity for Pearson test of English academic

Zheng, Ying (Pearson Language Tests, United Kingdom)*
De jong, John (Pearson Language Tests, United Kingdom)
Li, Jinshu (Pearson Language Tests, United Kingdom)

Abstract

The Pearson Test of English Academic (PTE Academic) is a new computer-based international English language test. It is designed to assess English language competence in the context of academic programmes of study around the world. To ensure that PTE Academic is valid and fit for its purpose, evidence was collected from the stages of test development through its implementation and launch. This study reports the comprehensive procedures, including two rounds of large-scale field tests and Beta tests that have been carried out to ensure the quality of the PTE Academic. The procedures that were adopted to ensure the validity and reliability of a complete machine scored test will be presented, i.e., the calibration of the automated scoring engines and the training of the scoring engines used to investigate the reliability of human ratings and machine-generated scores. To examine whether individual item performance on the Beta test was comparable to the field tests, the Delta method was used to identify items that were systematically more difficult for one set of examinees over another. Overall, the psychometric analysis of PTE Academic field test data and Beta test data accomplished the following: helped determine scoring models, provided data to be used for training and validating intelligent automated scoring systems, identified items with substandard quality, which were then eliminated, established how item scores can be combined to generate reporting scores, established the minimum number of items for each item type in the test structure, defined difficulty parameters for all items.

C07 08:30-10:00 Room211

C07-1

Structural modeling of TIMSS 2007 mathematics test data

Choi, Youn-jeng (University of Georgia, USA)*
Cohen, Allan S. (University of Georgia, USA)
Bandalos, Deborah L. (University of Georgia, USA)

Abstract

The TIMSS (Trends in International Mathematics and Science Study) testing program is designed to provide comparative information about student achievement in mathematics and science among participating countries. In this study, we focus on a structural analysis of Grade 4 mathematics with an eye to determining what student and country characteristics might be related to differences in test performance. This grade is important as it is typically when most countries begins formal education of rational numbers, particularly proportional reasoning. We focus on several sets of characteristics, including student, teacher, educational policy, and educational curriculum characteristics to try to determine which sets appear to have the most impact on student test performance. TIMSS data from the 39 countries participating in TIMSS 2007 Grade 4 mathematics will be used in this study. TIMSS 2007 used 14 test booklets administered randomly to students so we will use one test booklet among the 14 test booklets. Preliminary exploratory factor analyses from one randomly selected booklet (N = 13,446) indicates a single factor. Linear regression indicated several factors related to total math score over the 26 items in the booklet: public expenditure on education, pupil-teacher ratio, emphasis on a national curriculum, and provision of remedial instruction. Multilevel structural equation modeling will be used to further analyze these data for the conference paper.

C07-2

Presenting comparative cross national achievement tests results: Are league tables useful?

Griffin, Patrick (University of Melbourne, Australia)*
Care, Esther (University of Melbourne, Australia)
Ross, Kenneth (University of Melbourne, Australia)

Abstract

Large-scale cross-national studies of achievement often focus on differences between countries. Countries are typically ranked on the basis of average scores while taking into account standard errors of measurement. Data from the Southern and Eastern African Consortium for Monitoring Education Quality (SACMEQ) are reported here to demonstrate that there are more informative ways of reporting cross-national studies. The project assessed reading comprehension and mathematics among Grade 6 students in 15 sub-Saharan countries. It also addressed school resources, teaching strategies, and home and school influences on achievement. Reports from the project assume that competence and developmental levels of learning are more important and more informative than average scores. Also, the methods of reporting between-country differences developed within the SACMEQ project by Ross and others combine scores, variance and relationship to SES and equity measures. Using item response modelling, a series of competence levels in reading and mathematics are derived, and reports generated for the participating countries on the distribution of students at each level of competence. In a unique approach, teachers in many of the 15 countries were also tested for reading comprehension and mathematics competence. Teachers' and students' competence levels were mapped onto the same scales. This enables direct comparisons of teacher competence and student development. The results enable systems of education to be given advice on the formulation of policy for intervention. The intervention recommendations target the level of competence of both teachers and students.

C07-3

Do our item statistics harm the instructional sensitivity of achievement measures?

Kingston, Neal (University of Kansas, USA)*

Abstract

In this study 200 students will each be provided with one of two forms of a 15-item pre-test (randomly assigned), and then a lesson on a history standard on which they are unlikely to have received instruction in their school (this assumption will be checked as part of a questionnaire). After the lesson they will be tested again using both forms of the test with each student receiving all 30 items. Two weeks later they will be tested again using the same 30 items as in the immediate post-test. Separately, a group of 10 teachers will be asked to rate each of the 30 items on their expectation regarding the impact instruction would have on student performance on the item (instructional sensitivity). Also, three experts in social studies curriculum and assessment will rate the quality of the items for measuring achievement of the unit of instruction. For each item the following statistics will be calculated: item-total correlations based on the data from the pretest administration, increase in normalized item based on the data from the pretest and first posttest, increase in normalized item based on the data from the pretest and second posttest, average teacher judgment regarding instructional sensitivity, average expert judgment regarding item quality. The relationship among the five statistics for each item will be studied. This will include looking at the 15-item tests that would be created from the 30-item pool based on each approach.

C08 08:30-10:00 LT4

C08-1

Psychological testing status and challenges in a Southern African context

Foxcroft, Cheryl (Nelson Mandela Metropolitan University, South Africa)*
Mzwinila, Kefentse (Gaberone, Botswana, South Africa)
Janik, Manfred (Psychology Department, University of Namibia, Namibia)

Abstract

The purpose of this paper is to provide an overview of the status of psychological testing in three Southern African countries, namely, Botswana, Namibia and South Africa. In the process the challenges faced will be outlined and ways in which psychologists, Psychological Associations and universities are trying to address these challenges will be highlighted. The possibility of creating an African network of psychological associations will also be explored as one way of fostering greater collaboration and collective problem-solving regarding similar testing-related issues.

C08-2

The South African Personality Inventory: A progress report

Meiring, Deon (University of Pretoria, South Africa)
Van De Vijver, Fons (Tilburg University, The Netherlands & North-West University (Potchefstroom Campus), South Africa)*
De Bruin, Deon (University of Johannesburg, South Africa)
Marais, Carin (University of Johannesburg, South Africa)
Oosthuizen, Talitha (North-West University)

Abstract

Most of the personality inventories employed in South Africa are imported from Western countries and usually administered in English. These measures show important limitations, such as the use of English-language instruments among testees with an insufficient mastery of English, especially for the majority black groups in South Africa, and cultural bias. We provide a progress report of a project, aimed at developing a culture-informed, psychometrically sound inventory, called the South African Personality Inventory. The inventory will be based locally and derived from indigenous conceptions of personality in all 11 official South African languages. The presentation consists of three parts. The first briefly describes the theoretical background and the cultural context of the study as well as the two stages of the project: a qualitative stage to identify indigenous personality dimensions and a quantitative stage in which an inventory based on these dimensions is validated. The second presents the findings of the qualitative stage. The process is explained of how the original more than 52,000 utterances were reduced in a number of steps to a nine-cluster structure. We also explain how in the test development stage, 2,194 content-representative items were developed on the basis of the qualitative analyses. The third part of the presentation presents data of one of the ongoing psychometric validation studies. We report on the results of an experimental inventory for the Relationship Harmony facet that consisted of four subclusters that was administered to a multicultural sample group. We describe the analytical procedures and present preliminary findings.

C08-3

The use and impact of two languages on examination performance and implications for language policy in educational assessment: A sub-Saharan African case study

Yu, Guoxing (University of Bristol, United Kingdom)*
Rea-Dickins, Pauline (University of Bristol/Institute of Educational Development, East Africa)
Abeid, Mohammed (The State University of Zanzibar, Tanzania)

Abstract

A significant number of children around the world demonstrate their learning in formal examinations through English as a foreign or second language, and the function of English for subject learning and assessment has become increasingly controversial in the context of the UNESCO Education for All Agenda. For example, what impact does an unfamiliar language have in determining learning progression and outcomes and, ultimately, the educational quality of the system itself? In many sub-Saharan African contexts, over 50% of children leaving school are labelled as unsuccessful on the basis of their performance on examinations administered in English, which is neither their nor their teachers' first language. We report on the findings from two empirical studies of the SPINE research project (Student Performance in National Examinations; ESRC/DfID Major Research Grant RES-167-25-0263; www.bristol.ac.uk/spine) that is investigating the dynamics of language in learning and on the fair and ethical assessment of examination performance at the end of Basic Education in Zanzibar. This paper focuses on language issues in examinations and will report on: (i) the process of how students work through two languages (English and Kiswahili) in responding to examination questions in Mathematics, Biology and Chemistry, (ii) the differential effects on students' performance (n=800) in responding to three different versions of a Maths, Chemistry and Biology examination: English only, Kiswahili only, and English/Kiswahili bilingual versions of the three subjects, (iii) the relationships between these students' English language proficiency (in particular, in terms of lexical knowledge) and their demonstration of subject knowledge, and (iv) such relationships as evidenced in the national data of the Form II examinations in 2007-2009 (c. 70,000 students).

C08-4

The impacts of participant personalities on the judgments in modified angoff standard settings

Zhang, Sheng (Faculty of Education, Beijing Normal University, China)*
Zhang, Danhui (Faculty of Education,Beijing Normal University, China)
Tang, Keran (National Center of Educational Assessment, China)
Chi, Zhaoyan (National Center of Educational Assessment, China)

Abstract

The procedure of standard setting is to collect the judgments of qualified participants with various experience and expertise about the level of knowledge and skills required for a person to be considered as above the cut-score. Many studies have investigated objectivity of judgment and the extent to which one's final judgment is influenced by the background of participants. This study investigated the impact of participant personalities on judgments. A modified Angoff method with four iterative group discussions was employed in setting performance standards for the Chinese National Assessment of Education Quality in 2008. Influential data were also provided during the group discussion to modify participants' judgment. The personalities of each participant were also measured with the Big Five Personality Scale for the purpose of measuring participants' characteristics on agreeableness, conscientiousness, openness, neuroticism, and extraversion. It was found that people with different personalities showed distinct behavioral tendencies in this setting. Those with neuroticism tended to adjust their judgments based on the provided information, but they were not easily influenced by the other participants; those higher on agreeableness were more likely to agree with others, as well as reach consensus; those higher on openness were generally active in

the group discussion. Despite the discrepancies of person behaviors during the procedure, the findings suggested that the participant personality was not a significant factor in influencing the overall final standard setting judgment.

C08-5

Towards a holistic resiliency-to-depression model in Puerto Rican populations: Translation, adaptation, validation, and psychometric properties of spirituality and well-being instruments

Scharron-del Rio, Maria (Assistant Professor of Brooklyn College - City University of New York, USA)
Hill, Jill (Teachers College, Columbia University, USA)*

Abstract

Vulnerability-to-depression models for Puerto Ricans have included cognitive, interpersonal, and environmental factors (Bernal & Bonilla, 2003). While useful within a clinical setting, this model fails to incorporate an important cultural dimension: spirituality. Furthermore, it provides limited information regarding strengths and resiliency to this disorder. As a first step to assessing depression in Puerto Ricans from a holistic resiliency perspective, various spirituality and well-being instruments were translated (inverse translation method; Bravo et al., 1991; Brislin, 1970), adapted, and validated for this population. The convenience sample consisted of 286 undergraduate students (mean age = 21.21, 73.0% women) from the University of Puerto Rico, Río Piedras campus. Self-report instruments were used to assess various dimensions of spirituality (Spiritual Assessment Inventory, SAI; Spiritual Well-Being Scale, SWBS; Spiritual Trascendence Index, STI), depressive symptomatology (Beck Depression Inventory – Spanish, BDI-S; Symptom Check List – 36, SCL-36; Center for Epidemiological Studies – Depression scale, CES-D), happiness (Subjective Happiness Scale, SHS), optimism (Life Orientation Test- Revised, LOT-R), quality-of-life (Quality-of-Life Index - Spanish version, QLI-Sp; World Health Organization Quality-of-Life Scale – Brief version, WHOQOL-BREF) and social support (Cuestionario de Apoyo Social, CAS). Descriptive, reliability, factor analyses, and correlations were performed for the all instruments. A multiple regression analysis with physical quality-of-life, interpersonal quality-of-life, social support, optimism, and perceived relationship with a Higher Power as predictors of depressive symptomatology explained 61.4% of the variance (R2 = .614, p < .001). Implications for the multidimensional assessment of mental health and well-being of Spanish-speaking populations will be discussed.

C09 08:30-10:00 Room201

C09-1

The need for a new mode of test administration for the ITC guidelines

Bartram, Dave (SHL Group Ltd, United Kingdom)*

Abstract

When the ITC Guidelines on computer-based and internet delivered testing were published, we identified four different modes of administration: Open, Controlled, Supervised and Managed. The first two of these described modes where there was no supervisor or proctor present. In the case of the Controlled mode, there was control over the candidate's login, while for the Open mode anyone might be able to access the test. Developments over the last few years means that we now need to reconsider this typology as we see increasingly sophisticated methods being used to remotely supervise test candidates. Some of these are illustrated and their implications are discussed. In particular I will consider the degree to which we may now be able to better control test conditions and test taking behaviour remotely than was possible with traditional supervised group administration of paper and pencil tests. It would seem that degree of control and supervision is no longer simply correlated with the presence or absence of a test session supervisor.

C09-2

Worse than plagiarism? Firstness claims & dismissive reviews

Phelps, Richard P. (District of Columbia Public Schools, USA)*

Abstract

With a firstness claim, a researcher insists that s/he is the first to study a topic. With a dismissive literature review, a researcher assures the reader that no one else has conducted a study on a topic. Both are commonplace in testing research. Research literature reviews are not analytically taxing, but a thorough review requires a substantial amount of time. Meanwhile, scholars receive little credit for a thorough literature review, gaining much more for "original work". In "publish or perish" environments, lit reviews are impediments to progress. How prevalent are they? The number of hits on Google for certain phrases suggests the scale of the problem: "absence of research" ~57,000,000 "absence of studies" ~28,900,000 "this is the first study" ~9,830,000 "little research" ~2,620,000 "paucity of studies" ~2,560,000. Firstness claims and dismissive reviews may cause more harm than plagiarism, and may be more difficult to prevent. Dismissive reviewers not only steal credit and misrepresent the originality of their work; they suppress the dissemination of others' work. But, whereas plagiarists are more likely to be punished, as their intent is easy to establish, dismissive reviewers can fall back on the "honest mistake" excuse. As a result, society loses information, and the remainder is skewed in favor of those better able to publicize their work. Also, funders pay again for research that has already been done. The author will hypothesize the origins of the problem and recommend policies for mitigating it.

C09-3

The implications of systematic training for the validity of online ability testing

Preuss, Achim (cut-e GmbH, Germany)*
Wehrmaker, Maike (cut-e GmbH, Germany)

Abstract

Fluid intelligence is critical for a wide variety of cognitive tasks, and it is considered one of the most important factors in learning. Hence, fluid intelligence is a vital construct in aptitude testing. There is a long history of the assumption that fluid intelligence must be simply genetically determined. However, recent studies (see Jeaggi, Buschkuehl, Jonides

& Perrig, 2008) have indicated that fluid intelligence can be improved with training. If training improves fluid intelligence, it needs a measure for current psychometric fitness level to determine the true potential of a candidate in aptitude testing. Cognitive training data of participants (N=12,548) of a long-term online study is used to identify cognitive aspects that are subject to improvement by training and to aptitude testing. The results indicate that taking into account the cognitive training level is a vital element of test fairness for aptitude testing in general and for online pre-screening in particular.

C09-4

Unproctored testing in a virtual world: Does it matter?

Serlie, Alec W. (Erasmus University Rotterdam / GITP, Netherlands)*
Oostrom, Janneke K. (Erasmus University Rotterdam, Netherlands)
Born, Marise Ph. (Erasmus University Rotterdam, Netherlands)

Abstract

In the modern internet era, more and more people are being assessed by completing questionnaires via the world wide web. Quite often these questionnaires are completed under unproctored circumstances (Nye, Do, Drasgow, & Fine, 2008). Even though there are many advantages associated with this type of testing (Ployhart, Weekley, Holtz, & Kemp, 2003), one should also take possible disadvantages into account, such as faking, or impression management (Converse, Peterson, & Griffith, 2009). In the present study the score differences on personality measures in either proctored or unproctored circumstances were studied. Our hypothesis was that there would be a significant difference between the two circumstances, leading to higher personality scores in the unproctored situation especially when the stakes were high (assessment for selection) versus low (vocational assessment). In this field study, a group of 1,241 test-takers who had completed a FFM personality questionnaire in an unproctored situation were matched for age, gender, educational level and assessment type (selection vs vocational) with a group of 1,176 test-takers who had completed the questionnaire in a proctored situation. The results showed that the test-takers who had completed the personality questionnaire in an unproctored situation had significantly higher scores on Extraversion, Openness to experience and Self-presentation than the test-takers who had completed the personality questionnaire in a proctored situation. There was also a significant interaction with assessment type, in that the scores were even higher in the high-stakes selection situation. One of the main implications of this study is that care should be taken when interpreting personality scores acquired in a unproctored situation, especially when the stakes are high.

C09-5

Unproctored internet testing: Faster, better and cheaper? Choose two!

Sharf, James (Sharf & Associates, Employment Risk Advisors, Inc., USA)*

Abstract

The digital age is not about technology, it is about information – the content of tests and assessment instruments – intellectual property which is in play with a scanner and the click of a mouse. Internationally, testing practices tend to be influenced more by the presence of multinational corporations and consulting firms than by legal pressures (with the exception of United states, Canada, and South Africa). In the U.S., legal compliance concerns shape testing practices. Most countries, however, do not rely on validity evidence to refute claims of employment discrimination. For many test developers, the unproctored internet testing (UIT) train has left the station. More than two thirds of employers who test are already engaging in UIT - a substantial increase from 31% in 2005. To date, because UIT has not been directly evaluated legally in the United States, there is little guidance for the international test provider other than following the ANSI (2007) for certification programs and the ITC (2006) guidelines for Internet-based testing. International copyright, trademark, trade secret and patent laws will be discussed, and detailed. “Proprietary IP Paranoia” describes the real world experience of test developers such as DDI’s Bill Byham and Hogan Assessment Systems’ Bob Hogan. The moral of their IP and UIT experience; With thanks to Bob Hogan, “Constantly innovate so that your material is always more fresh than that which has been stolen from you.”

C10 10:30-12:00 Room201

C10-1

Item selection rules in computerized adaptive testing: Is there any way to know which is the best one?

Barrada, Juan Ramon (Universidad Autonoma de Barcelona, Spain)*
Olea, Julio (Universidad Autonoma de Madrid, Spain)
Ponsoda, Vicente (Universidad Autonoma de Madrid, Spain)
Abad, Francisco José (Universidad Autonoma de Madrid, Spain)

Abstract

We will define four different objectives that should be achieved with the item selection rule chosen for a computerized adaptive testing program, with their more common indicators: a) to maximaze reliability (indicator: RMSE); b) to maximaze item bank security (indicator: test overlap); c) to meet the test constraints in content (indicator: number of examinees receiving an exam not meeting the constraints); and d) to reduce the cost of the testing program maintenance (indicator: mean discrimination parameter of the administered items). Usually, only one or two objectives are taken into account when comparing item selection rules. Also, the common studies show the relation between objectives in a very limited fashion. We will show how the preferred item selection rule is highly dependent on the objectives included in our analysis and that, when doing an extensive manipulation of scenarios, our understanding of item performance is improved. We will illustrate this point with two examples. In the first one, we will show how the number of strata used in the alpha-stratified method depends on whether we only consider reliability and security as objectives (one stratum) or if we also incorporate the cost to replace items (more than one stratum). In the second example, we will show how an extensive manipulation of the values of the maximum exposure rate allowed for the items shows that the alpha-stratified method does not offer better reliability given the same level of test security, when compared with other item selection rules.

C10-2

Impact of compromised items on item calibration

Han, Kyung (Chris) T. (Graduate Management Admission Council, USA)*
Guo, Fanmin (Graduate Management Admission Council, USA)

Abstract

When test items are compromised over time due to overexposure and/or security breaches, item parameter drift (IPD) becomes a serious threat to test validity and fairness. Extensive research has been conducted to investigate the impact of IPD on item parameter and proficiency estimates using simulated and real data. Most of these simulation studies, however, overly simplified IPD situations in which the item parameters drift with all test takers. In reality, test items are exposed only to a small percentage of test takers, and so the IPD would occur only with those examinees. This study employed simulation studies to examine the impact of IPD on item calibration when items were exposed only to a portion of test takers under two different test situations: fixed test forms and adaptive testing. The study hypothesized that the item parameter estimates and model fit would start to be influenced by IPD as the percentage of test takers exposed to those items increased. The findings of this study provide practitioners with insight into efficient ways to handle item exposure while minimizing possible item compromise.

C10-3

Assessment of differential item functioning with Raju's area method and the DIF-free-then-DIF strategy

Lin, Jyun-Ji (National Chung Cheng University, Taiwan)*
Su, Ya-Hui (National Chung Cheng University, Taiwan)
Wang, Wen-Chung (The Hong Kong Institute of Education, Hong Kong SAR, China)

Abstract

Differential item functioning (DIF) analysis is a standard practice in many large-scale testing programs to ensure test fairness. Raju (1998, 1990) proposed the area DIF detection method, which is based on quantifying the gap between item characteristic curves of reference and focal groups. DIF assessment relies on a clean matching variable. Scale purification procedures have been developed to purify the matching variable. Unfortunately, if a test has as many as 20% or more DIF items, the test will be seriously contaminated and the subsequent DIF assessment is misleading. Wang (2008) thus proposed the DIF-free-then-DIF strategy to diminish the impact of DIF detection when tests contain DIF items. The purpose of my study is to apply this strategy to Raju's area method and assess its performance in a series of simulations. Five independent variables were manipulated in the simulation study: (a) method (the standard area method, the area method with scale purification, the area method with a clean anchor by design, and the area method with the DIF-free-then-DIF strategy); (b) test length; (c) percentage of DIF items in the test; (d) mean ability difference between groups; (e) anchor length. The results showed that when there were DIF items in the test, the area method with a clean anchor by design performed the best, the standard area method performed the worst, and the area method with scale purification and the area method with the DIF-free-then-DIF strategy performed similarly. In conclusion, the DIF-free-then-DIF strategy is applicable for the area method.

C10-4

Multidimensional model for mixed response items: A comparison of CFA and IRT approaches

Liu, Hongyun (School of Psychology, Beijing Normal University, China)*
Luo, Fang (School of Psychology, Beijing Normal University, China)

Abstract

In the present paper, the Mixed (dichotomous and polytomous) item response factor analysis model is analyzed and compared in two approaches, which are: the item response theory (IRT) and the structural equation model approach. Two Monte-Carlo simulation studies are conducted to examine the effect factors of parameter estimations in unidimensional and multidimensional situations, respectively. The Monte Carlo simulation studies revealed: (1) Similar parameter estimates were obtained from the SEM and IRT parameterizations. Even with a small sample and the IRT estimates converted to SEM parameters, the MWLSc, and MML/EM results were strikingly similar. But with a small sample size and a long test, WLSc did not obtain the convergence parameter estimations, although in a short test where WLSc estimates were obtained, the estimates are consistently more discrepant than those produced by the other estimation techniques; (2) The precision of the estimators is enhanced as the quantity of the sample increases; (3) The precision of item factor loading and of item difficulty parameters is influenced by the test length; (4) In whole, the precision of item parameter estimates in an SEM framework is higher than that of the IRT framework. Both structural equation modeling (SEM) and item response theory (IRT) can be used for factor analysis of dichotomous item responses. In this case, the measurement models of both approaches are formally equivalent. They were refined within and across different disciplines, and make complementary contributions to central measurement problems encountered in almost all empirical social science research fields.

C10-5

Establishing effect size guidelines for interpreting the results of differential bundle functioning analyses using SIBTEST

Walker, Cindy M. (University of Wisconsin, Milwaukee, USA)*
Zhang, Bo (University of Wisconsin, Milwaukee, USA)

Abstract

The issue of fairness in educational testing has long been a concern. This issue is mainly addressed through Differential Item Functioning (DIF) analyses, which determine whether equal ability examinees from distinct groups have equal chance of obtaining a correct response (Holland & Wainer, 1993). DIF studies for single item performance are frequently exploratory, hence, they provide little information on why DIF has occurred. An alternative approach is to conduct Differential Bundle Functioning (DBF) analyses based on the performance of a bundle of items (Douglas, Roussos, & Stout, 1996). As these items are bundled meaningfully in terms of cultural background or cognitive domains, substantive hypotheses can be determined a priori as to why items are functioning differentially. However, while guidelines exist for interpreting the magnitude of single-item DIF (see Roussos & Stout, 1996b), such guidelines have not been proposed for interpreting item-bundle DBF. This study was set to fill that gap. To achieve that, a simulation study

was designed to examine the properties of the bUNI statistic obtained from SIBTEST. In the simulation, important test characteristics, such as bundle size, DIF magnitude, sample size, and impact, were studied. Major findings include a) the bUNI parameter estimate was approximately normally distributed for all conditions explored, and b) the average of the DBF statistic increased by a multiplicative factor as the number of items in the bundle increased. Based on these findings and the guideline for interpreting single item DIF, a guideline for interpretation DBF was derived.

C10-6

An approach to implement adaptive testing using Item Response Theory

Natarajan, Venkatesa (MeritTrac services Pvt.Ltd., India)*
Padaki, Madan (MeritTrac services Pvt.Ltd., India)
Bhardwaj, Shipra (MeritTrac services Pvt.Ltd., India)

Abstract

In India, as most of the large scale testing is conducted in the paper-pencil (offline) mode, it is important to arrive at models of implementing IRT in an offline/paper-pencil mode. MeritTrac has experimented in conducting an IRT-based test in a paper-pencil mode for the analytical abilities test for engineering graduates. With the help of item characteristics calculated prior to the test, a 6-item test with increasing item difficulty was created as a test form on paper. Normally, research shows that a 6/10 item test can be compared to 25 or more items in the test. The test was then administered to the candidates in an offline mode. The responses of the test taker were then entered in Test taker [student] Tracking software that had been specially coded for this purpose. The output of this gives an estimation of test taker's true score as if he has taken the parent 25 item tests. Since it is not very feasible to conduct an online test everywhere especially in a country like India, the importance of Adaptive Testing in offline mode increases many folds.

C11 13:30-15:00 Room211

C11-1

Assessing culturally specific and invariant factors for the international foundations of medicine examination

De Champlain, Andre F. (National Board of Medical Examiners, USA)
Gessaroli, Marc (National Board of Medical Examiners, USA)*

Abstract

The mobility of physicians across borders for employment and training is commonplace, as evidenced by a number of government sponsored initiatives, including the Bologna Process in Europe. However, few assessment tools are available to provide a common measure of core medical knowledge for health professionals seeking employment or training abroad. To address this shortcoming, the National Board of Medical Examiners developed the multilingual International Foundations of Medicine (IFOM) examination with a number of European partners. The purpose of this study was to apply a bottom-up methodology proposed by Gessaroli & De Champlain (2009) for test form assembly. This methodology, based on Hui & Triandis' (1985) cross-cultural framework, could be useful to construct equivalent IFOM test forms composed not only of measurement invariant or ETIC items, but also culturally specific (unidimensional) or EMIC items. Our study was based on the responses of 1861 examinees who either completed the Italian (N=1413) or Portuguese (N=448) version of the 200-item 2009 IFOM. The first phase of our analysis resulted in ultimately identifying 23 ETIC items. Using these items as common anchors across both IFOM versions, we were also able to identify additional (unidimensional) EMIC items, scaled to the same metric as ETIC items, for inclusion into each respective form. We believe that our findings provide evidence to support the bottom-up approach proposed for the construction of equivalent test forms across cultures. The implications of our results will be further discussed within the framework of the test adaption literature.

C11-2

A synthetic framework for creating comparable test forms across populations that contain both common and culturally specific test items

Gessaroli, Marc (National Board of Medical Examiners, USA)*
De Champlain, Andre (National Board of Medical Examiners, USA)

Abstract

Establishing construct and item/measurement equivalence is crucial when adapting forms for use across different cultures. The first step in establishing this equivalence is usually an assessment of construct equivalence. When existing tests are adapted for use in a different language or culture the complexity of the content and its possible differences across cultures often results in a failure to establish construct equivalence. Without construct equivalence any assertions of item/measurement equivalence are tenuous. The purpose of this paper is to outline a methodology that addresses the problem of construct and measurement/item equivalence simultaneously. The method has two very general steps In the first step, using all the items in the test as the initial item pool, a subtest that is common to all cultural groups is built in a systematic way such that it is unidimensional and measurement invariant. The second step identifies additional items within each cultural group that measure the common construct. The result is a different test for each cultural group which has a common set of items with measurement invariance that allows for a common scale across the different groups and unique items for each cultural group that reflect the relevance of the item in that group in measuring the common construct. Integrated into the methodology at various stages of the process are statistical tests for the important psychometric properties such as unidimensionality, measurement invariance, etc. The strengths and weaknesses of such an approach will be discussed.

C11-3

Multidimensional Rasch and CFA analysis of Chinese version of Professional Learning Community Scale (PLC-CV)

Lee, John Chi-kin (Department of Curriculum and Instruction, Facuty of Education, The Chinese University of Hong Kong, Hong Kong SAR, China)*
Zhang, Zhonghua (Department of Educational Psychology, Faculty of Education, The Chinese University of Hong Kong, Hong Kong SAR, China)
Song, Huan (Center for Teacher Education Research, Faculty of Education, Beijing Normal University, China)

Abstract

This study used the multidimensional Rasch model and confirmatory factor analysis (CFA) to examine the construct validity and detect the substantial differential item functioning (DIF) of the Chinese version of professional learning community scale (PLC-CV). A total of 1670 primary school teachers were administered the PLC-CV which was composed of 28 items. Partial credit model was suggested to have a better goodness of fit than that of the rating scale model. DIF based on multidimensional Rasch model found eight items having substantial differential functioning for different groups. The psychometric quality of the revised PLC-CV which was composed of the remaining non-DIF 19 items was re-examined using multidimensional Rasch model and CFA. The correlations between the subscales indicated that the factor shared sense of purpose and focus and the factor staff support and cooperation were highly correlated. Finally, the item difficulty and step parameters of the 19 items and the reliabilities of the four subscales were displayed. Based on these findings, the directions for the future research were discussed.

C11-4

In search of qualified accountants: Validation of an accounting quality scale using the Rasch model

Mamauag, Maria Felicitas (Marife) M. (De La Salle-College of Saint Benilde, Philippines)*
Magno, Carlo (De La Salle University, Philippines)

Abstract

The study validated a non-cognitive measure called the "Accounting Quality Scale" (AQuaS). The AQuaS measures an individual's likelihood of possessing personal dispositions considered desirable for good accountants across four subscales namely: Planning and thinking, human relations, business experience, and lean culture. The subscale of the AQuaS is composed of 60 items reflecting the examinees' attitude towards needed functional competencies of accountancy students. Two forms of the AQuaS were pre-tested among 300 accountancy students. Item review was established through expert judgments. Cronbach's alpha of both forms was at least .90, indicating high internal consistency among the items. Through a factor analysis, items with factor loadings of .4 and above were initially accepted and 4 factors were extracted. From the 240 original items, there were 120 good ones that comprise the final form of the AQuaS. The items for each factor were further calibrated using the one-parameter Rasch model, and those that show adequate fit based on MNSQ and Z-values were accepted in the final form.. The AQuaS may be further examined with its relationships with other measures of personal attributes, considering that the manifestations of the factors of the AQuaS are clearly evident depending on the examinees' dispositions.

C11-5

The equivalence of the Spanish version of the American board of family medicine in-training examination

O'Neill, Thomas (American Board of Family Medicine, USA)*
Royal, Kenneth (American Board of Family Medicine, USA)
Raddatz, Mikaela (Xavier University, USA)

Abstract

Family medicine residents of American Board of Family Medicine (ABFM)-accredited residency programs are eligible to sit for the In-Training Examination (ITE), which is a precursor to the board's Certification/Recertification Examination. The ITE specifications mirror those of the core portion of the board certification examination. Typically, the ITE is administered in English but in 2008 and 2009 a translated Spanish version was administered in Quito, Ecuador. Of note are concerns that translating the test items into Spanish may have changed the meaning of the questions and, consequently, the meaning of correctly answering them. Additionally, residency training in Quito might be substantively different than residency training in the United States. Because the hierarchy of the items by difficulty is essentially a description of the construct being measured, the question arises, is the hierarchy of the items by difficulty—the construct—the same across language versions; In order for the measures to be comparable, it is assumed that the same construct underlies both examinations. This assumption is testable given that the same items, although translated, were on both forms (Spanish and English) of the test. The items were independently calibrated using each cohort with a dichotomous one-parameter logistic model (Rasch, 1960). The calibrations were then plotted and compared. The two sets of calibrations (Spanish and English) were positively correlated, r (238) = 0.66, $p < 0.01$, and when plotted, were evenly distributed about the identity line. This suggests that the same construct is in effect in at least a very general way. Due to the small sample size of the Spanish cohort, significant differences between calibrations were unlikely to be detected. Although the study suggests that the same general construct is in place across both language versions, some subtle and, perhaps, not so subtle differences may have gone undetected due to power issues. The data from 2008 is included herein; the data from 2009 will be available for the conference presentation.

C11-6

Setting minimal English proficiency standard for entry-level healthcare professionals: A comparison of standard setting methods

Woo, Ada (National Council of State Boards of Nursing, USA)*
Marks, Casey (National Council of State Boards of Nursing, USA)

Abstract

One of the core components for obtaining healthcare licenses in the United States is passing the required licensure examinations. While licensure examinations measure knowledge in the discipline of interest, they do not consider additional factors that may affect a licensee's ability to practice safely and effectively, such as language proficiency. Globalization of the workforce has resulted in increased foreign test-taker volume for US licensure examinations. For example, the number of National Council Nursing Licensure Examinations administered to internationally-educated registered nurse candidates rose from 30,575 in 2003 to 51,381 in 2008. As the number of internationally-trained healthcare professionals seeking licensure in the US increases, so does the need to establish language proficiency prerequisite for granting licenses to practice. Using the nursing profession as a case study, the present study will discuss the processes by which the National Council of State Boards of Nursing conducted English proficiency standard setting workshops using the TOEFL CBT, TOEFL iBT, IELTS and PTE Academic. Different standard setting methodologies were used in these workshops due to differences in item formats and test design across these language proficiency tests. This study will discuss efficacy of the Simulated Minimally Competent Candidate paradigm, Examinee Paper Selection method and the Modified Angoff method as used in language proficiency standing setting for healthcare professionals. Subject matter experts' selection criteria, standard setting training and discussion of the minimally competent candidate will also be addressed in the present study.

C12 13:30-15:00 Room201

C12-1

Test use in Norway: The problem of adapting imported tests to small language communities

Egeland, Jens (Vestfold Mental Health Care Trust, Norway)*
Høstmælingen, Andreas (The Norwegian Psychological Association, Norway)

Abstract

High quality tests are important for correct descriptions of function or for diagnostic purposes. Particularly in small language societies, there is a risk that many tests are not satisfactory adapted when imported or translated. In Norway, as probably in other countries, the scope of the problem is not fully recognized as the testing practices is mostly unknown. As part of a new initiative to improve test quality, the board of test and testing practices of the Norwegian Psychological Association surveyed the test use of Norwegian psychologists. About 1/5 of all Norwegian psychologists answered a comprehensive questionnaire regarding test use and test practice. Thirty-one tests of cognitive function were used by a least 10% of all psychologists, or psychologists within one clinical discipline. Correspondingly, 32 different clinical inventories exceeded the 10. percent limit for common use. All but 4 of the total number of instruments had been translated from English, or developed within an Anglo-American culture. Few of the tests had been formally standardized or normed in Norway. Implications: There is a need for developing low cost adaption studies of translated tests that do not depend on a commercial marked. Examples of such studies are discussed.

C12-2

Cultural validity and item response with indigenous samples: Implications for the MMPI-2 and MMPI-2-RF

Hill, Jill (Teachers College, Columbia University, USA)*
Scharron-del Rio, Maria (Brooklyn College, City University of New York, USA)
Skolnik, Avy (Teachers College, Columbia University, USA)

Abstract

The MMPI-2 is the most widely used and researched personality assessment instrument in the world (Dana, 2000; Greene, 2000). Despite this status, only four published studies exist which examine the validity of the MMPI-2 for use with Native American adults (Greene et al., 2003; Hill et al., in press; Pace et al., 2006; Robin et al., 2003). Given the scarcity of research on the MMPI-2 with Native Americans, questions remain about its continued wide use in Indian Country (Robin et al., 2003). For example, can the MMPI-2 and its derivatives be culturally validated for use with Indigenous groups in the United States; What are the steps involved in this type of cultural validation process? What are the ethics of using such a tool to measure personality and psychopathology with American Indian adults when there is still limited information about the meaning of individual responses to items as well as insufficient understanding regarding significant scale elevations based on cultural factors rather than pathology? This presentation examines item-level MMPI-2 data from a Southwestern Plains Nation and an Eastern Woodland Nation, both of which are located in Oklahoma. Item-level similarities and differences between the two Indigenous samples are addressed on scales L, F, 1, 4, 6, 8, and 9. Differences in item endorsement, comparing the two Indigenous samples with the MMPI-2 normative sample, are explored from a cultural-contextual perspective. Implications for the most recent version of the instrument, the MMPI-2-RF, will also be addressed.

C12-3

Adaptation of Children's Depression Inventory on Russian adolescent sample

Belova, Alexandra (Psychological Institute of Russian Academy of Education, Russia)
Malykh, Sergey (Psychological Institute of Russian Academy of Education, Russia)*

Abstract

An adaptation and slight modification of the Children's Depression Inventory (CDI; Kovacs, M., 1992) on a Russian nonclinical adolescent sample was made. The sample consisted of 713 teenagers (335 boys and 378 girls) aged 13-17. Among them were 208 younger teenagers aged 13-14 and 505 elder teenagers aged 15-17. All respondents answered CDI questionnaire by themselves in a range of other emotion related questionnaires. Besides the adaptation we made a slight modification of the CDI in order to increase measure consistency: one item about high irritability (based on literature data and good consistency with other data) was added and one question about fear that something bad will happen was removed as showed low consistency with rest of the questions. Modified CDI questionnaire (27 questions) enjoyed quite high internal consistency (Alpha = 0.874). Based on Maximum likelihood factor analysis results 6 factors were revealed:

Negative mood, Negative self-esteem, Anhedonia and social isolation, Somatic concerns, Externalizing and Interpersonal problems. Though all factors account for about only 30% of data dispersion which corresponds to some other results of CDI factor analysis (e.g., Carey M.P. et al. Children's Depression Inventory: Construct and Discriminant Validity Across Clinical and Nonreferred (Control) Populations// Journal of Consulting and Clinical Psychology 1987, 55, 5, 755-761.).

C12-4

Factor structure of PTSD symptoms among Filipinos trauma survivors

Mordeno, Imelu (University of St. Joseph, Macau, China)*

Abstract

Several studies have used confirmatory factor analysis (CFA) to evaluate the latent structure of traumatic distress symptoms as measured by the Posttraumatic Stress Disorder Checklist – Civilian Version (PCLC–C; Weathers, Litz, Herman, Husk & Keane, 1993) among various trauma-exposed populations. However, these models were not yet tested in the Philippine context. The present study used CFA to evaluate seven models, including intercorrelated and hierarchical versions of two models with the most empirical support. Data were derived from a heterogeneous sample of trauma-exposed Filipinos. A model of PTSD with an intercorrelated four-factor model separating avoidance and numbing symptoms into distinct factors was found superior compared to different competing models. The research further related the four factors to other constructs to lend more validity to the model.

C12-5

Discovering psychometric properties of psychological capital questionnaire

Yan, Gonggu (School of Psychology, Beijing Normal University, China)*
Lu, Xinxin (China National Offshore Oil Company, China)

Abstract

Psychological capital is a new construct coined by positive psychologists in last decade in the domain of human resource management and development. To further the research and application it is critical to measure the construct adequately under investigation. This study examines the reliability and validity of a Chinese-translated version of psychological capital questionnaire (CPQ), designed by Dr. Luthans and explores the effect of age and work experiences on each of the subscales. The sample consisted of 262 offshore oil workers and their supervisor from 7 platforms in China who work at the platform on the sea for 28 days at a time. The internal structure of the scales and their inter-relatedness were evaluated. Results indicated that the reliability of the subscales for self-efficacy, hope, and resilience was comparable to those in previous studies, but not for optimism. Criterion-related validity defined as the correlation between psychological capital index and task performance and engagement rated by their supervisor were r=0.21 and r=.27 respectively.

C13 15:15-16:45 LT3

C13-1

Evaluation of potential for creativity in children

Barbot, Baptiste (Child study center, Yale University, USA)*
Besançon, Maud (University of Paris Descartes, France)
Lubart, Todd (University of Paris Descartes, France)

Abstract

Since Guilford (1950), creativity is an essential concept viewed as ability normally distributed which can be developed. Thus, an important issue for both creativity research and practical applications is to assess creativity to support and guide creativity development. However, existing creativity assessment tools are limited, especially for their lack of up-to date norms and theoretical framework. We present a new instrument that allows creative potential to be measured (Evaluation of Potential Creativity, EPoC, 2010), based on solid theoretical and research-based evidence for creativity assessment. EPoC is a modular, domain-specific tool, which presently includes verbal and graphic subtests that measure the two key modes of creative cognition—divergent-exploratory thinking and convergent-integrative thinking—in elementary and middle-school students. Psychometric results concerning the instrument are presented (CFA and external validation with other cognitive and personality measures), as well as an original, internet-based scoring system that enhances inter-rater reliability. The instrument, developed initially with a sample of 300 French school children, is currently adapted into several languages. EPoC can be used as an efficient diagnostic tool to identify creative potential as well as Creative giftedness, and to monitor progress, using pre-tests and post-tests, in educational programs designed to enhance creativity.

C13-2

Development and analysis of PrepTeST, a cognitive screening tool for preparatory level children

Care, Esther (University of Melbourne, Australia)*
Griffin, Patrick (University of Melbourne, Australia)
Lo, Rebecca (University of Melbourne, Australia)
Kan, Mimi (University of Melbourne, Australia)

Abstract

A cognitive screening tool – the Preschool Teachers Screening Tool (PrepTeST) – was developed in order to provide teachers with a resource to help identify five-year old children of non-English speaking backgrounds who might need additional learning support. N = 140 children completed the tool, administered by probationary psychologists on a one-to-one basis, as well as the Naglieri Nonverbal Ability Test (NNAT) and Wechsler Individual Achievement Test (WIAT-II). The children were identified from non-English and English speaking backgrounds respectively; within the non-English speaking background group, major subgroups comprised Middle Eastern, Asian and Eastern European children. Data from the NNAT and WIAT are presented in order to examine evidence of construct validity of the developed tool. PrepTeST items which appear to discriminate between non-English and English speaking background students are identified and discussed in order to clarify item types that may discriminate between groups. Also, the difficulty of appropriately attributing group differences to ethnic background or to socio-economic status is highlighted. The implications of this difficulty is discussed in the context of accurate identification of between group differences. PrepTeST has the potential to provide preschool teachers with an accessible tool to identify student needs in a timely and efficient manner.

C13-3

Applying the fuzzy measurement on situational judgment tests: Comparisons of the fuzzy linguistic variables and fuzzy composite scores

Chiou, Hawjeng (Department of Business Administration, National Central University, Taiwan)*
Ou, Tsung-Lin (Department of Business Administration, National Central University, Taiwan)

Abstract

During the past decade, Situational Judgment Tests (SJTs) have recently enjoyed a resurgence in attention in the research literature. However, one critical issue remaining unsolved is the selection of scoring methods of SJT. It is due to the feature of SJT in which the items often do not have objectively correct answers. In addition, the examinees have difficulties to pick an exact answer from the options. Based on the viewpoints of Fuzzy Composite Score (FCS) and fuzzy linguistic variables, the aim of this research is to deal with judgment uncertainty which was ignored on the issue of scoring SJT. The two proposed methods can provide fuzzy space for linguistic judgment and allow researchers to compare the difference of SJT item score among various degrees of linguistic vagueness. Take the forced-choice scale of SJT as an example, this study simulated sample responses for 5 various degrees of vagueness to calculate SJT item score and compare differences between the fuzzy scoring and crisp scoring. Analysis of results concludes that: (1) as degree of linguistic vagueness increases, the divergence in SJT samples' score was reduced. (2) Through modifying degree of linguistic vagueness, proposed two fuzzy methods can increase the discrimination between subjects to cut down the error rate when screening applicants. Finally, practical and academic viewpoints about fuzzy scoring on SJT were discussed for the future study.

C13-4

A comparative study of attribute mastery between regional entities within the U.S.: An analysis of TIMSS 2007 using cognitive diagnostic modeling

Lee, Young-Sun (Teachers College, Columbia University, USA)*
Park, Yoon Soo (Teachers College, Columbia University, USA)
Taylan, Didem (University of Missouri - Columbia, USA)

Abstract

Mathematics educators and psychometricians often wonder how best to estimate the mastery of grade-appropriate curricular skills and are also interested in examining regional differences in attribute mastery within the same country. In the 2007 administration of the 4th grade mathematics Trend in International Mathematics and Science Study (TIMSS), two U.S. states—Massachusetts and Minnesota—were included as benchmarking participants, which are regional entities that follow the same assessment procedures as the countries. They ranked 6 and 7, respectively, among 56 total participants, which were both above the U.S. national sample that ranked 14th. This framework and design of the TIMSS assessment provides an ideal structure to conduct an empirical analysis of attribute mastery within the U.S. Although various attempts have been made to make inferences on attribute mastery based on information from student performance based on broad content domains, traditional methods of analyzing large-scale assessments such as classical test theory or item response theory have shown to be ineffective to provide attribute-level information. In contrast, employing a cognitive diagnostic modeling (CDM) approach, attribute mastery as well as cognitive diagnostic information can be attained. This study shows that there is more information to be gained using a CDM framework that can be translated directly for classroom application. Results showed that even with high performance, examinees from the two states lacked mastery of key attributes. Furthermore, the data revealed attributes where students lacked further practice. Advantages of the CDM approach are discussed as well as CDM-based method to filter distractor response categories.

C13-5

Identifying the domains of scientific thinking

Magno, Carlo (De La Salle Universoty, Manila, Philippines)*

Abstract

The present study further explains the nature of scientific thinking by exploring and confirming its factors. There is a need to rethink the components of scientific thinking because of the lack of integration and coherence of domains in the composition of its construct. A total of 240 items were constructed referring to characteristics of scientific thought and it was administered to 528 college students taking a science course and who are currently in the stage of writing their thesis. The underlying factors of the 240 items were identified using a principal components analysis with varimax rotation. Analysis of the scree plot showed that four factors can largely explain the total variance (10.94%). The grouping of the items were reviewed and they were labeled as practical inclination, analytical interest, intellectual independence, and discourse assertiveness. These new set of factors were administered to a similar sample (N=1000) and the factors were confirmed in a measurement model. A four factor-model of scientific thinking was structured and tested using Confirmatory Factor Analysis. The results showed that the four factors of scientific thinking significantly increase with each other.

C13-6

CTT, IRT, and G-DINA analysis of TIMSS 2007

Park, Yoon Soo (Teachers College, Columbia University, USA)*
Lee, Young-Sun (Teachers College, Columbia University, USA)

Abstract

Mathematics educators and psychometricians often wonder how best to estimate the mastery of grade-appropriate curricular skills and are also interested in examining regional differences in attribute mastery within the same country. In the 2007 administration of the 8th grade mathematics Trend in International Mathematics and Science Study (TIMSS), two U.S. states—Massachusetts and Minnesota—and three Canadian provinces—

Québec, Ontario, and British Columbia—were included as benchmarking participants, which are regional entities that follow the same assessment procedures as the countries. They ranked 6, 7, 8, 9, and 13, respectively, among 56 total participants, which were all above the U.S. national sample that ranked 14th. This framework and design of the TIMSS assessment provides an ideal structure to conduct an empirical analysis of attribute mastery within the U.S. and Canada. Although various attempts have been made to make inferences on attribute mastery based on information from student performance based on broad content domains, traditional methods of analyzing large-scale assessments such as classical test theory, item response theory, or generalizability theory have shown to be ineffective to provide attribute-level information. In contrast, employing a cognitive diagnostic modeling approach, attribute mastery as well as cognitive diagnostic information can be attained. This study examines fine-grained attribute mastery using the deterministic, inputs, "and" gate (DINA; Junker & Sijtsma, 2001) model for each regional entity that can be used as diagnostic information for educational instructors to improve student performance and learning in mathematics education.

C14 15:15-16:45 Room211

C14-1

Applicability of an Implicit Association Test (IAT) for measuring implicit risk-taking

Bittner, Jenny V. (Jacobs Center on Lifelong Learning, Jacobs University Bremen, Germany)*
Johannes Becker (Kassel University, Germany)
Tanja Neumann (Kassel University, Germany)

Abstract

Implicit attitudes are judgments about persons, objects, or actions that are stored in memory and can be activated without cognitive effort by situational cues (Wilson, Lindsey & Schooler, 2000). Implicit attitudes are typically assessed with measurement techniques that measure the strength of associations existing in memory, e.g. with the Implicit Association Test (IAT). It is expected that an IAT makes it more difficult for the tested person to manipulate test results than in self-reports, like interviews or questionnaires. The goal of the present study was to investigate the applicability of an IAT for the prediction of criminal offenses. Therefore, an IAT was compared to a questionnaire in their practicability to measure risk-taking. The sample were offenders in German prisons and their risk attitudes were compared to a control group. It was found that offenders differed in their implicit as well as explicit risk attitudes from the control group, as they were significantly more risk-taking. In addition, those offenders that had committed crimes in an impulsive, unintentional manner showed significantly higher implicit attitudes on the risk IAT than those offenders which had committed planned, reflective crimes. These results demonstrate that for the prediction of criminal offenses it would be helpful to measure both, implicit attitudes with reaction times and explicit attitudes with self-report, since their results may differ for reflective versus impulsive information processing.

C14-2

The development of a bilingual adolescent adjustment inventory

Atluri, Ashok (Andhra University, India)
Kosuri, Madhu (Andhra University, India)
Dasari, Venkata Venu Gopal (Andhra University, India)*

Abstract

Recognizing the need for a standardized adjustment inventory for adolescents in Telugu (3rd most widely spoken language in India) we developed a bilingual (Telugu and English) inventory that measures five major adjustment domains of adolescent life, namely Home, Health, Academic, Emotional and Social adjustments. An initial pool of 334 English items were translated into Telugu and back translated into English. Item readability was established on a sample of 435 6th grade children with an inclusion criterion of having ≥ 75% of the sample agreeing that they understood the meaning of the item. Item level equivalence between the English and the Telugu versions was established statistically on a separate sample of 132 bilingual adolescents. Psychometric analysis of the inventory, conducted on a cross sectional sample of 2073 adolescents (Boys=1050; Girls=1023), involved item discrimination, item-total correlation and reliability analysis. Items with a significant correlation of ≥ .30 with the corresponding dimension were retained resulting in a final set of 192 items. Cronbach's alpha for the five scales ranged between .76 and .86 and the split half reliability ranged between .69 and .86. This inventory would be useful for identifying the counseling needs of adolescents and devising effective intervention strategies.

C14-3

Psychological Teatment Inventory (PTI): A new measure for planning treatment and assessing psychotherapy outcome

Giannini, Marco (Department of Psychology, University of Florence, Italy)
Gori, Alessio (Department of Psychology, University of Florence, Italy)*

Abstract

This study describes a new instrument for evaluating interventions and outcomes using repeated measurements. The Psychological Treatment Inventory (PTI) includes interacting biological, psychological, behavioral, social, and environmental factors across various domains central to planning psychological treatment and evaluating its outcome: a) Psychological Resources; b) Quality of Life; c) Psychological Types; d) Symptomatology; e) Attachment Styles; f) Predominant Defense Styles; g) Treatment Outcome Predictors; h) Negative Treatment Indicators. Methods. In order to verify the psychometric properties of the PTI a series of Exploratory Factor Analysis (EFA) and Confirmatory Factor Analyses (CFA) have been performed. Internal consistency of each scale has been verified using the Cronbach's alpha coefficient. Some aspects of concurrent validity have been verified using the Persons correlation coefficient (r). Discriminant validity has been verified with a series of ANOVA between a clinical sample and a non-clinical sample. Exploratory Factor Analysis (EFA) showed the dimensionality of each scale with good values of internal consistency. Results were confirmed by Confirmatory Factor Analysis (CFA) for each cluster. For what concerns the aspects of

Concurrent Validity the PTI showed good correlation with some of the most self-report measures used for the assessment of psychopathology. Thanks to its good psychometrics properties the PTI can be used for the repeated measurement of client status over the course of therapy and at termination for both, research and practice. Clinicians should then be able to choose the best therapeutic approach and intervention strategies, thanks to an efficient procedure for analysis of the information obtained at intake.

C14-4

Development and validation of a questionnaire to measure the service needs of families with children with developmental disabilities

Leung, Cynthia (The Hong Kong Polytechnic University, Hong Kong SAR, China)*
Lau, Joseph (Child Assessment Service, Department of Health, Hong Kong SAR Government, Hong Kong SAR, China)
Chan, Grace (Child Assessment Service, Department of Health, Hong Kong SAR Government, Hong Kong SAR, China)
Lau, Beverley (Child Assessment Service, Department of Health, Hong Kong SAR Government, Hong Kong SAR, China)
Chui, Mandy (Child Assessment Service, Department of Health, Hong Kong SAR Government, Hong Kong SAR, China)

Abstract

The aim of this project was to develop and validate a service needs questionnaire (SNQ) on the service needs of families with children with developmental disabilities. The SNQ and a measure of parenting stress were administered to 105 parents of children diagnosed with learning/behaviour problems and 233 parents of children attending primary schools. Initial Rasch analysis results indicated inadequate distinction of the categories and the fit statistics of three items were outside the acceptable range. The categories were collapsed and the removal of two misfitting items resulted in a scale which conformed to the Rasch expectations. For validity, the scale correlated positively with parenting stress, and it could differentiate between parents of children diagnosed with learning/behaviour problems and those attending primary schools. The internal consistency estimate (Cronbach Alpha) was above .70. The SNQ could be used to help identify the needs of families with children with developmental disabilities.

C14-5

Towards the development of an indigenous inventory to assess cognition and behavior of chinese gamblers: An exploratory study.

Tao, Vivienne Y. K. (University of Macau, Macau, China)*
Wu, Anise M. S. (University of Macau, Macau, China)
Cheung, Shu Fai (University of Macau, Macau, China)
Tong, Kowk-kit (Unviersity of Macau, Macau, China)

Abstract

The prevalence of gambling is found to be constantly high among Chinese and Chinese ethic communities in the West. However, most research on Chinese gamblers directly used the translated measurements developed from Western samples. It is questionable to apply those scales and generalize results to Chinese gamblers. In addition, most gambling studies were carried for clinical purposes and have over-focused on problem gambling which constitutes a tiny portion of the gambling population (e.g., less than 2.5% in Macao, Fong & Bozorio, 2005). The majority of the non-problem gamblers have been rarely investigated. To fill the void, the current research aimed to develop an indigenous instrument to measure cognitions and behaviors of the general Chinese gamblers. A telephone survey was conducted in which nearly 800 gamblers were randomly sampled and successfully interviewed. Exploratory factor analysis was performed to identify factors of the cognitive and behavioral domains. Findings of the factor structures, their implications and the future directions of research will be discussed.

C14-6

Development and adaptation of indigenous anger expression inventory based on State Trait Anger Expression Inventory, Spielberger, 1988) , in Urdu, in Pakistani cultural context and psychometric properties of the Pakistani questionnaire

Shahid, Mamoona (Government M. A. O. College, Lahore, Pakistan)*
Najam, Najma (International Islamic University, Korakoram, Gilgit, Pakistan)

Abstract

Anger assessment has got increasing attention of psychologists due to its great role in developing hypertension and heart diseases. In the present research State Trait Anger Expression Inventory by Spielberger (1988), a 44-item scale was adapted into Urdu because of unfamiliarity with the English language. No such standardized anger scale was available in Urdu in Pakistan. The adaptation and translation procedure was completed in 4 phases. In the first phase, 44 sentences of STAXI were translated into Urdu with the help of three standardized dictionaries. The two choices of Urdu translation, closest to actual meanings were selected and given to 5 patients and 5 experts for the selection of most appropriate translation. The same procedure was adapted in four phases till a final list of 44 Urdu sentences was prepared from the most frequently selected sentences by patients and experts. In the 4th phase the same list was translated back into English by three experts and inter-rater concordance was found in positive direction. Psychometric properties and of the adapted version were also checked. State and trait anger contain 10 items each, and anger expression includes 14 items. A sample of 56 male and female hypertensive patients between ages 35-65 years was taken from different hospitals. Internal consistency and reliability analyses showed sufficient alphas (i.e. internal consistency) ranges from .61 to .77 and test-retest coefficients (i.e. stability) for the adapted version. The correlations of the six subscales with criterion variable (Urdu adaptation of STAXI) were found in the expected direction.

C16 15:15-16:45 Room201

C16-1

Applying mixture IRT models to TIMSS data

Alexeev, Natalia (University of Georgia, USA)*
Cohen, Allan (University of Georgia, USA)
Templin, Jonathan (University of Georgia, USA)

Abstract

The Trends in International Mathematics and Science Study (TIMSS) is an international assessment to measure trends in mathematics and science learning. The aim of TIMSS is to improve the teaching and learning of mathematics and science by providing data about students' achievement relative to different curricula and instructional practices. A recent development in TIMSS research is application of mixture IRT models, particularly mixture Rasch models (MRM), to investigate qualitative and quantitative differences among latent groups in the examinee population. The use of the MRM is prevalent in such studies in part due to its simplicity as well as to the availability of software packages for estimating model parameters. Some previous research suggests, however, that using reduced IRT models, such as the Rasch model, may lead to over-extraction of latent classes and, therefore, to misinterpretation of results. The purpose of this study is to investigate how mis-specified mixture IRT models can affect extraction of latent classes and thus interpretation of the data. The problem and solutions will be illustrated using TIMSS 2007 Mathematics Test for Grade 8.

C16-2

An investigation of measurement equivalence of dual-language accommodation science test for bilingual learners

Ong, Saw Lan (Malaysia Science University, Malaysia)*

Abstract

Measurement equivalence is important when accommodation is being given in testing. Evidence of comparability is necessary to ensure valid interpretations made based on scores derived from administration of accommodated test. The purpose of dual-language accommodation is designed to allow students to access the content measured in a science test. It is hope that the use of dual-language test enables accurate measure of student knowledge and skills in science content area. The purpose of this study is to compare the measurement equivalence of the accommodated dual-language science test and the English-only science test. The English-only science items were hypothesized to be multidimensional, whereas the dual-language science items were hypothesized to be unidimensional with the elimination of the language factor. The equivalence is evaluated with three types of empirical analyses. First, descriptive statistics are examined to make a comparability decision at either the test or item level. A number of indicators such as the mean total scores, reliability index at test level, and the item difficulties, item discrimination for item level are compared. Second, dimensionality analyses are carried out to assess the dimensional structure of the dual-language science test. If the dimensional structure of the accommodated test is different, this provide evidence that the dual-language science test is measuring something different from the English-only version. Third, differential item functioning is analyzed to assess the construct equivalence of the test.

C16-3

The effect of standardized testing on student achievement: Meta-analyses and research summary

Phelps, Richard P. (District of Columbia Public Schools, USA)*

Abstract

This twelve-year study summarizes the research literature on the effect of standardized testing on student achievement. It reviews several hundred quantitative, qualitative, and survey studies conducted from the early 20th century on. I employed two search methods—keyword searches and citation chains. Reviewed studies are of three methodology types: quantitative (N ≈ 250), revealing about 600 measurable effects; studies that measured perceptions of effects through surveys (N ≈ 250), producing about 800 effect measures; and qualitative studies revealing about 300 measures. Over two thousand other studies were reviewed and found to be unsuitable for inclusion. For quantitative and survey studies effect sizes are calculated directly and summarized with weighted and unweighted means. Qualitative studies are classified into positive and negative effect categories. Mean effect sizes for quantitative studies are moderately positive (> 0.5 & 1.0). Among qualitative studies, 95 percent found that learning had improved after the introduction of a test or an increase in the stakes of a test. These results contradict the claim popular with many U.S. education researchers that this research literature does not exist.

C16-4

Combining Bayesian networks with two-tier items to modeling students' learning bugs and sub-skills

Shih, Shu-Chuan (Department of Mathematics Education, National Taichung University, Taiwan)*
Kuo, Bor-Chen (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)
Yang, Chih-Wei (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)

Abstract

The main purposes of this study are to develop the two-tier mathematics diagnostic tests based on Bayesian networks and explore the efficiency of combining Bayesian networks with two-tier items for modeling students' learning bugs and sub-skills in time calculation after students have learned the related contents. Six steps are involved in this study: developing the student model based on Bayesian networks that can describe the relations between bugs and sub-skills; constructing two-tier items that students can be provided an opportunity to reveal their bugs and sub-skills in time calculation contents; using two-tier items for evidence model creation and completing the cognitive diagnostic model that combining Bayesian networks with two-tier items; administering test items for sixth graders in elementary school; estimating the network parameters using the training sample and applying the generated networks to bugs and sub-skills diagnosis using the testing sample; and assessing the effectiveness of the combined models work in predicting the existence of bugs and sub-skills. The tests are administered to 300 sixth grade students. The responses of 240 samples are used as a training data set for building Bayesian networks and others are treated as a testing data set for evaluating the holdout classification accuracies of the combined model. The results show that using combined model to diagnose the existence of bugs and sub-skills in individual students can get good performance.

C16-5

Mathematics test equating under the Graded Response Model (GRM)

Syaifuddin, M. (Faculty of Teacher Training and Education, University of Muhammadiyah Malang, Indonesia)*
Purwono, Urip (FAculty of Psychology, Universitas Padjadjaran, Bandung, Indonesia)

Abstract

This study investigates (1) the best method, (2) the minimum number of anchor items required, (3) the effect of test length and sample size, and (4) the effect of ability difference on equating mathematical tests. Under a "common item nonequivalent groups design", simulated and real data were studied. For the simulation study, number of items, sample sizes, and ability distribution were varied. Fifty replications were carried out. The PARSCALE V3.2, EQUATE V2.1, and mean & sigma (MS) was employed to estimate the item parameters and calculate the equating coefficients respectively. The RMSD was used as criterion of the accuracy of the equating. The results show that: (1) Except for sample size of 800 with 20 items, Stocking & Lord (LS) method is superior to the MS. For these conditions, the quality of the equating under MS3 and MS4, and SL gave the same results, and both are better than the MS2 method; (2) the number of anchor items as well as the test length and sample size has an effect on the quality of the equating; (3) the equating was more accurate for groups with the same ability distribution than those with different ability distribution. It was concluded that the Stocking & Lord method has a merit as a procedure of choice in equating a mathematical test.

C15 16:45-18:15 LT3

C15-1

Verbal Comprehension Index (VCI) subtests of Wechsler Intelligence Scale for Children, fourth edition (WISC-IV UK): Adaptation and norms development in Pakistan

Ambreen, Saima (National Institute of Psychology, Quaid-i-Azam University, Islamabad, Pakistan)*
Kamal, Anila (National Institute of Psychology, Quaid-i-Azam University, Islamabad, Pakistan)

Abstract

The research was aimed for the adaptation and norms development of the Verbal Comprehension Index (VCI) subtests of WISC-IV UK (Wechsler, 2004) in Pakistan. The research was completed through three studies. Study-I was concerned with the adaptation of VCI subtests. Initially problems in the items of original VCI subtests regarding the difficulty level, understanding and cultural relevance were identified. Then five items in vocabulary subtest and four items in information subtests were replaced with expert opinion. Afterwards, the item functioning and psychometric strength of the adapted Verbal Comprehension Index (VCI-P) subtests was assessed. The alpha coefficient for VCI composite was found to be quite satisfactory. Study-II was concerned with the establishment of reliability and validity evidence for the VCI-P. The stability coefficients ranged from .82 (word reasoning) to .92 (vocabulary), while alpha reliability coefficients ranged from .72 (comprehension) to .88 (vocabulary). For structural validity, inter-subtests correlations (ranging from .63 to .78) and subtest-VCI correlations (ranging from .84 to .94) were computed. Factorial validity resulted in existence of a single factor. In study-III age based standard scores (scaled and composite scores), percentile and test-age equivalent norms were developed for VCI-P in Pakistan on a sample of 801 students (boys = 385 & girls = 416). All norms were quite comparable with the norms of the original version. Mean differences in VCI Equivalent scores showed an increase of 15- 20 points when computed and analyzed with Pakistani norms as compared with the UK norms.

C15-2

A cross-cultural investigation of reasoning ability – identifying factors for differential facet functioning

Bertling, Jonas Pablo (University of Münster, Germany)*
Freund, Philipp Alexander (University of Osnabrück, Germany)
Holling, Heinz (University of Münster, Germany)
Kuhn, Jörg-Tobias (University of Münster, Germany)

Abstract

In a globalized world with multicultural societies, differential item functioning (DIF) constitutes a severe threat to test-fairness and standardized assessments in educational and selection settings. Rule-based item generation represents a feasible approach to investigating cultural fairness on the level of item facets. The differential facet functioning (DFF) model, a particular version of the Generalized Linear Mixed Model (GLMM), helps to explain DIF effects substantively by detecting bias directly at the level of item facets. We examine DFF in the Latin Square Task (LST) based on samples of 454 German and 203 Russian university students. The LST is a rationally constructed measure of reasoning ability that resembles the worldwide popular Sudoku puzzles. Our results extend findings of previous studies by showing how basic cognitive operations contribute to reasoning across two ethnically diverse samples. Furthermore, by modelling a number of nested models, we compare the differential impact of cultural background as well as prior Sudoku experience on the functioning of cognitive task parameters.

C15-3

Testing invariance of the measurement model underlying standardized spatial ability test across three areas in Taiwan

Jeng, Hi-Lian (National Taiwan University of Science and Technology, Taiwan)*
Chen, Yu-Fang (National Taiwan University of Science and Technology, Taiwan)

Abstract

The purpose of the study is to test for measurement invariance of the Standardized Spatial Ability Test developed by Taiwan's College Entrance Examination Center (SPAT-CEEC). With stratified proportional sampling, the final number of sample was 1641 after listwise deletion, and was classified into Northern (n=695), Central (n=421), and Southern (n=525) areas. The theoretical structure of the analytical procedures started from a suggested version of SPAT, which is an 18-item, two-factor structural model from an earlier study, to proceed with the following model comparisons. Analyses of multiple samples factorial invariance in structural equation modeling were applied to explore the questions: 1. Is SPAT invariant across three areas? 2. Is the latent factorial structure of SPAT the same across three areas? 3. Is the latent means structure of SPAT the same across three areas? The results showed that all of the model, factor loadings, measurement errors, factorial structure variances and covariances, and the latent means structure of SPAT were the same across three areas; the test is measurement invariant and equivalent in the Northern, Central, and Southern areas. It is also interesting to find that the latent means are different in the same order as Northern>Central> Southern for both spatial factors.

C15-4

Assessing variations of item fit and difficulty of a problem solving test for children in a one-parameter Rasch model across language and number of operations

Magno, Carlo (De La Salle University, Manila, Philippines)*
Ong, Paul Kelvin (De La Salle University, Manila, Philippines)
Liao, Vernice (De La Salle University, Manila, Philippines)
Alimon, Rosee (De La Salle University, Manila, Philippines)

Abstract

An eight item mathematical problem solving test was constructed for grade 6 Filipino students following standard content and skills. The content of the problem solving test cover skills on addition and subtraction of fractions, decimals, and whole numbers. The test has both Filipino and English language versions (where items are parallel). Both of these versions have forms with items requiring an answer for a single operation and double operation. The items were reviewed by teachers in mathematics for grade 6. When the items were calibrated using a one-parameter Rasch model, higher person (.92-.97) and item reliabilities (.48-.60) were obtained as compared to Cronbach's alpha (.40-.72). The difficult items (positive logits) outnumber the easy items (negative logits) for the English versions and involving multiple operations. However, items in English language had better fit (MNSQ within .8 to 1.2) than the ones in Filipino language. When each test type was correlated with the MSLQ scores, the math test in Filipino language involving single operation had a significant correlation coefficient. The results of the study imply the (1) responsiveness of even minimal items using the IRT approach; (2) difficulty of mathematics problems include multiple operations; (3) feasibility of the English language to solve mathematics problems; and (4) facilitative aspect of the Filipino language in the use of learning strategies in solving mathematics problems.

C15-5

The drawing of neckers cube as an initial cognitive screening of immigrants/refugees participating in language training programs

Sundseth, Øyvind (NPF, Norwegian Psychological Association, Norway)
Marberger, Tove Kanestrøm (NPF, Norwegian Psychological Association, Norway)*

Abstract

We suspected that unidentified cognitive problems could be a major obstacle for adequate second language aquisition in adult immigrants/ refugees, participating in second language training programs. Our traditional test/screening procedures are known to be insufficient/ not adapted for use with persons with a non-western background and insufficient school experiences. We wanted to establish the usefulness of a low cost/effort instrument without cultural bias. It is assumed that important verbal and visual cognitive processes are needed to attain the skills that is sufficient to copy a task as for instance the Neckers cube. This is a two - dimensional ambiguous figure - giving a three- dimensional impression/mental representation. It is hypothesised that the copying task demands important visuo-constructive, executive and possible working memory capacities. Theese skills are commonly thought to be important during written language aquisition. We wanted to identify the potential for learning and at the same time avoid time consuming and high effort standard cognitive testprosedyres. The goals are to identify learning potentials in immigrants/refugees participating in language training and avoid comprehensive cognitive testing prosedyres with potential ethical and cultural biases/problems. Our findings underline the importance of early identification of possible cognitive obstacles to learning a second language. Identifying such obstacles for learning could be an important factor for sucsessful second language pedagogical program for immigrants and refugees.

C17 16:45-18:15 Room201

C17-1

Scoring multidimensional pairwise preference tests with item response theory: Comparisons with other methods and formats

Stark, Stephen (University of South Florida, USA)
Chernyshenko, Oleksandr (Nanayng Technological University, Singapore)*
Ho, Moon-Ho Ringo (Nanayng Technological University, Singapore)

Abstract

Despite compelling evidence showing validity of personality measures in many contexts, concerns about faking have motivated researchers to explore alternatives to the traditional single stimulus format for administering items. Our particular interest is in the multidimensional pairwise preference (MDPP) format which, in comparison to multidimensional tetrads and pentads, appears to be less strenuous for examinee to take and is more psychometrically tractable. In this paper we will focus on our item response theory (IRT) based approach to construction and scoring tests comprised of multidimensional pairwise preference items (e.g., Stark & Drasgow, 2002). We will present results of a recent empirical study comparing scores derived from single stimulus and multidimensional pairwise preference items corresponded in tests of high dimensionality. We will also explore the use of confirmatory factor analysis (Thurstonian approach) for analyzing the MDPP data, and compare with the results (e.g., parameter estimates, latent scores) obtained via the IRT method.

C17-2

A novel parameterization for Partial Credit Model with numerous categorical responses and its estimation using WinBugs

Fung, Tze-ho (Hong Kong Examinations and Assessment Authority, Hong Kong SAR, China)*

Abstract

Item response theory (IRT) models have been widely adopted in education research and assessment. Amongst various models, the Partial Credit Model (PCM) is commonly employed in various areas. However, if the marks range of an item is too large (say from 0 to 20), the commonly available software for PCM (e.g. WINSTEPS and ltm R package) may not be able to estimate such a large number of parameters. In this study, we follow the similar ideas from RUMM2020; but re-formulate the PCM model under different parameterization using at most 4 parameters for each item, even though marks assigned to each of them could spread over a wide range. Such a formulation could largely reduce the number of parameters that have to be estimated. Compared with the formulation used in RUMM2020, the one proposed in the study is somehow more clear-cut and simple. Under the new formulation, the parameter estimation is conducted in Bayesian approach using WinBugs, which is a freeware for Bayesian analysis. Simulated and real-life data sets are used to test the applicability of the new formulation. We found that the parameter estimation under the new formulation could be achieved with reasonable accuracy in both simulated and real-life situations. The novel parameterization proposed in the study could widen the applicability of the PCM in educational assessment. Moreover, WinBugs is a freeware whose uses are without restrictions. Thus, any interested parties could employ the proposed approach in their research studies and applications.

C17-3

Discussion of the three-parameter logistic model: What is the best guess at the guessing parameter?

Han, Kyung (Chris) T. (Graduate Management Admission Council, USA)*

Abstract

Ever since Birnbaum (1968) introduced the three-parameter logistic model (3PLM), several studies have pointed out technical and theoretical issues regarding c-parameter and its interpretation (Lord, 1974, 1975, 1980; Kolen, 1981; Holland, 1990; Hambleton, Swaminathan, & Rogers, 1991). It is surprising, however, that those studies have had little impact on the current use of 3PLM in the field. Unfortunately, for example, it is often observed that imprudently interpreting c-parameter as a guessing parameter causes critical problems in test construction and standard setting. This study attempted to introduce a logical argument for reconceptualizing the guessing and the problem-solving processes and to suggest an alternative model to 3PLM. Examples and discussions in this article revisit the implications of a-, b-, and c-parameters of 3PLM and suggest practical solutions to avoid inappropriate use of item parameters of 3PLM.

C17-4

An item response model with hierarchical latent traits for polytomous items

Huang, Hung-Yu (Hsuan Chuang University, Taiwan)*
Wang, Wen-Chung (The Hong Kong Institute of Education, Hong Kong SAR, China)
Chen, Po-Hsi (National Taiwan Normal University, Taiwan)

Abstract

Many latent traits in the human sciences have a hierarchical structure. Huang, Wang, and Chen (2010) proposed the item response model with hierarchical latent traits for dichotomous items in which both the 2nd-order and the 1st-order latent traits are considered. In this study we further extend their work and develop a new model for polytomous items. To be general, the item response function in the new model can follow the generalized partial credit model (GPCM), the graded response model (GRM); the partial credit model (PCM), or the rating scale model (RSM). Data were generated from the new model with item response functions following the GPCM, GRM, PCM, and RSM and were analyzed with the true model using WinBUGS 1.4. Five test statistics based on the posterior predictive model checking (PPMC) were used to assess model-data fit, and the Pseudo-Bayes factor (PsBF) and Bayesian DIC index were computed for model comparison. Three 1st-order and one 2nd-order latent traits were assumed. The factor loadings of the three 1st-order latent traits were set as either diverse (.9, .6, and .3) or high (.9, .8, and .7). The data set contained responses of 1,000 persons to three tests, each test having 20 four-point items. Ten replications were conducted. The simulation results showed that all the model parameters can be recovered fairly well; the DIC is effective for model comparison, and the PPMC is helpful for the assessment of model-data fit. In conclusion, we have successfully developed a new item response model with hierarchical latent traits for polytomous items.

C17-5

A new approach of testing the Rasch model

Kubinger, Klaus D. (Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Austria)*
Rasch, Dieter (Department of Statistics, University of Agriculture Vienna, Austria)
Yanagida, Takuaya (Division of Psychological Assessment and Applied Psychometrics, University of Vienna, Austria)

Abstract

In correspondence with pertinent statistical tests, it is of practical importance to design data-sampling when the Rasch model is used for calibrating an achievement test. That is, determining the sample size according to a given type-I- and type-II-risk, and according to a certain effect of model misfit which is of practical relevance is of interest. However, pertinent Rasch model tests use chi-squared distributed test-statistics, whose degrees of freedom do not depend on the sample size or the number of testees, but only on the number of estimated parameters. We therefore suggest a new approach using an /F/-distributed statistic as applied within analysis of variance, where the sample size directly affects the degrees of freedom. The Rasch model's quality of specific objective

measurement is in accordance with no interaction effect in a specific analysis of variance design. The simulation study (100 000 runs for each of several special cases) proved that the nominal type-I-risk holds as long as there is no significant group effect. Analysing a certain DIF, this /F/-test has fair power, consistently higher than Andersen's test.

C17-6

Avoiding misuse of Rasch model in norm development

Yang, Zhiming (Educational Testing Service, USA)*
Liu, Ou Lydia (Educational Testing Service, USA)

Abstract

For most individually administered clinical assessments, a high quality norm, usually in a form of a raw total score to scale score conversion table, is of the utmost importance. Unlike large-scale assessments, most clinical assessments are not usable without a test norm. The Rasch model is frequently used to develop such a norm due to its convenience and powerful function. WINSTEPS, as one of widely used Rasch model software, is often used to generate a raw total score to proficiency (theta) conversion table, which typically serves as a test norm after linear transformation and hand-smoothing. This conversion table, however, can easily be misused in norm development. For example, when the vertical scaling analysis employs concurrent calibration method, this table is not correct and cannot be used as the foundation of norm tables. This study investigates some typical misuses of the Rasch model in the development of a test norm, such as misuse of the raw total score to proficiency conversion table, misuse of the estimates for extreme raw scores, and misuse of the estimates for those who have missing values, etc. We will conduct a simulation study to show the impact of the different misuses on test results. We will also provide an empirical example which illustrates the practice difference between misused norm tables and the correct table. We will conclude with recommendations for avoiding misuse of the Rasch model in norm development.

C18 08:30-10:00 Room211

C18-1

Prevalence of missing data in survey assessment

Dodeen, Hamzeh (United Arab Emirates University, UAE)*

Abstract

There is a general lack of attention to the problem of missing data in research in general and survey research in particular. Though critical, the problem is often inadvertently ignored. Missing data negatively affects the validity (internal and external) of the research, decreases the statistical power of tests, and makes the final sample unrepresentative of the actual population. This study aimed at determining the prevalence of missing data in real survey data and their effects on some statistical characteristics of the data. Of the 250 real survey data sets accessed from the internet, only 119 data sets were deemed appropriate to be analyzed. The analysis included determination of the percentage of missing values, comparison of the initial and final sample size, and gender differences with regard to missing data. Results indicated that missing data is indeed a widely prevalent problem in surveys. Of the 119 data sets analyzed, 49 surveys (41.2%) had more than half of their data missing and 54 surveys (45.4%) had less than 25% of missing data. On the whole, the average percentage of missing data was (38%). No significant gender differences were found, the average percentage of missing data for males being 37%, and that for females being 38%. Yet another significant finding was that, on average, the response rate of surveys was only 62% of the initial sample. In effect, missing data caused the male-female ratio to change from .91 to 1.08. The paper concludes with a discussion on the results and recommendations.

C18-2

Designing a test score report for multiple score users

De Jong, John H. A. L. (Pearson Language Tests, United Kingdom/ VU University Amsterdam, Netherlands)
Zheng, Ying (Pearson Language Tests, United Kingdom)*

Abstract

Recently the assessment literature has paid increased attention to the need for test score reports that provide information that is comprehensive and transparent for both candidates and score users. For example Hambleton at ITC 2008 and other meetings has pointed out that valid use of test scores depends on making score reports understandable and user-friendly. Communicating the score of a test, its meaning and how it should be used is not a simple matter and providing all information in a way which maximizes the probability that the interpretation by candidates and by other score users is valid sets high requirements because of the complexity of the matter to report and the variety of the target audiences. This contribution provides an example of score reporting consisting of a set of electronically linked documents that provide an immediately interpretable overview of scores and subscores in graphic and numeric formats and in addition provides full information on the meaning of the scores in terms of descriptors and information on the error of measurement. The possibilities for usage by different groups of stake holders will be discussed.

C18-3

An exploratory study of cognitive diagnostic model for error pattern attributes

Kuo, Bor-Chen (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)*
Yang, Chih-Wei (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)
Wu, Huey-Min (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)

Abstract

The purposes of this study are to develop mathematics diagnostic tests based on cognitive diagnostic models and explore the efficiency of modeling the error pattern attributes by the models. The mathematical abilities defined in the framework for national assessment of educational progress including the conceptual understanding, procedural knowledge and problem solving. Most research in cognitive diagnostic models focuses on estimating conceptual attributes that a student has or has not mastered. In this study, the procedural attributes and error pattern attributes are included in the cognitive diagnostic models to provide the teacher with information about instructional needs of groups of students.

C18-4

Value madded models and accountability

Swaminathan, Hariharan (University of Connecticut, USA)*
Rogers, H. Jane (University of Connecticut, USA)

Abstract

With the advent of the No Child Left Behind (NCLB) legislation, the issue of accountability has come to occupy center stage in public education. The policy governing Adequate Yearly Progress (AYP), designed to assess the effectiveness public education, requires the comparison of students' performance at a single point in time to pre-set standards. Recent research has documented the problems associated with and the inappropriateness of using the measure of academic growth as dictated by NCLB. On the other hand, the value added assessment procedures based on academic growth of students have been heralded as the tool for educational accountability and have been implemented in several states, notably Tennessee for documenting teacher effectiveness. While considerable research has been conducted on the usefulness of the AYP measure for accountability, the psychometric and statistical issues surrounding Value Added Models have been less well understood. The purpose of this paper is to examine the psychometric and statistical issues surrounding Value Added Assessment and Value Added Models. In particular, the growth model and projection model procedures developed by the authors for conducting value added assessment in a US state will be described and compared. In addition, the use of such models for identifying and providing resources for children and schools/districts at risk of failing AYP is described. Data from the state assessment will be used to illustrate the methods described in the paper.

C18-5

Adaptive personality assessment for unsupervised pre-screening

Preuss, Achim (cut-e GmbH, Germany)
Wehrmaker, Maike (cut-e GmbH, Germany)*

Abstract

Conventional Big Five orientated personality inventories (normative and ipsative) have proven their effectiveness for applicant screening under supervised conditions. However, these types of instruments are very prone to targeted biased responses which make them less effective in unsupervised conditions. Furthermore, in order to establish sufficient reliability these instruments are very time-consuming for participants which causes problems in acceptance under various circumstances. With adallocTM a new method is presented which enables an adaptive personality assessment online. It is demonstrated how the adallocTM method allows bias-robust assessment in a fraction of the test time of conventional methods. Results from an international deployment of the adallocTM method for unsupervised online pre-screening of undergraduates (N=95,314) are presented along with results from in vitro studies on the effect of targeted biased responses.

C18-6

Tamper-resistant aptitude pre-screening by real-time item generation

Preuss, Achim (cut-e GmbH, Germany)
Wehrmaker, Maike (cut-e GmbH, Germany)*

Abstract

The internet has offered new ways for deploying psychometrics for applicant pre-screening. In pre-screening settings, online assessment is typically and most effectively administered in an unsupervised mode, i.e., no test administrator is present while a candidate takes the online assessment. The online and unsupervised test administration settings call for quality criterion and technical solutions that go beyond those of classical testing. The extended quality criteria and technical requirements will be explained. Implementations and evaluations of this extended standard will be shown exemplarily that are in action for graduate selection (N= 35,325), for selection of technicians (N=7.963) and for selection of apprentices (N=192,431). Re-test results (supervised on-site) show that security issues, candidate authentication and tamper attempts of unsupervised test administrations can be successfully managed by real-time item generation.

C19-1

Testing the equivalence of self-concept configuration between Hong Kong and Singaporean chinese

Cheng, Christopher H. K. (City University of Hong Kong, Hong Kong SAR, China)*

Abstract

A prerequisite condition of cross group comparison on a "construct" is to ascertain its measurement equivalence. While self-conceptions are highly consonant with the social norms and cultural context, it can be predicted that the "self" configuration of the Singaporean and Hong Kong Chinese should be similar due to their cultural heritage. In view of the ethnic Chinese background of the two populations, an emically developed instrument, the Chinese Adolescent Self-Esteem Scales (CASES), was used. Initially, the psychometric properties of CASES were assessed in the two samples separately. In both samples the CASES were found to possess satisfactory measurement reliability, and could be fitted with a seven correlated factors model (general self-esteem and six domain-specific esteems, i.e., social, intellectual, appearance, moral, family, physical/sports). The focus of the present study was on the simultaneous testing of factorial invariance across the two populations. Specifically, three invariance models were tested, namely, M1 factor structure, M2 factor loadings, M3 inter-factor relations. It was found that all cross-sample invariant models were supported, suggesting that the self-concepts of Singaporean and Hong Kong Chinese are equivalent, from the most generic level (factor structure) to the structural configuration level (inter-factor structure). Having established the factorial equivalence, however, the author noted some between group differences, for example, Hong Kong people seemed to have lower physical self-concepts than Singaporean Chinese. As a concluding note, the author suggested the use of simultaneous testing of multiple groups SEM before comparing between group differences, especially when the construct is highly abstract.

C19-2

Measuring physical self concept with the psdq: An analysis of differential item functioning with German, Russian, and Turkish adolescents

Freund, Alexander (University of Osnabrück, Germany)*
Tietjens, Maike (University of Muenster, Germany)
Alfermann, Dorothee (University of Leipzig, Germany)
Aşçi, F. Hülya (Başkent University, Turkey)

Abstract

The Physical Self Description Questionnaire (PSDQ; Marsh, Richards, Johnson, Roche, & Tremayne, 1994) is a 70-item instrument designed for the measurement of 11 components of the (physical) self concept. Numerous studies have investigated the construct validity and generalizability to other cultural contexts, though mainly on the component level. Recently, Item Response Theory (IRT) methods (i.e., Samejima's (1969) Graded Response Model) have been applied as well, extending the analyses to the item level. In this study, we investigate the psychometric properties of the PSDQ on both the item and scale level by applying IRT methods of a different class [Rating Scale Model (Andrich, 1978), Partial Credit Model (Masters, 1982)] in a cross-cultural context to data obtained from German and Russian adolescents. Differential Item Functioning (DIF) describes the situation where, after controlling for the level of the latent propensity, the probability to respond in a certain category of the self-report rating scale is influenced also by group membership, so that the item "functions" differently for members of distinct groups. We analyze DIF in each of the 11 PSDQ scales in a Generalized Linear Mixed Model framework. By taking into consideration data from different cultural groups, we are able to make meaningful interpretations for the nature of DIF with respect to the cultural factors involved.

C19-3

Conceptual and methodological challenges in achievement motivation and self-construct research among very remote indigenous Australian students.

McInerney, Dennis (The Hong Kong Institute of Education, Hong Kong SAR, China)*

Abstract

This paper reports on a large scale quantitative psychometric study as well as interviews with school personnel, students, and elders of Aboriginal communities conducted in very remote Indigenous communities in the Northern Territory Australia. Thirteen school sites and 1044 participants contributed to the study. The study examines the inter-relationships between multiple achievement goals, self-concept, self-regulation, and learning processes and their relationship to achievement outcomes in communities that are very remote geographically and culturally from Western settings. The specific foci of this paper are: 1. To describe the validation processes adopted to maximize the cultural validity of multiple achievement goals and self-constructs used in the research; 2. To describe the use of confirmatory factor analysis and structural equation modelling with data obtained from very remote Indigenous participants; 3. To describe the theoretical, methodological, cultural and logistical difficulties experienced as the research was conducted and, 4. To address the critical issue whether Western oriented psychometric research can provide any real answers to the severe underachievement of remote Indigenous students in school settings.

C19-4

DECAS personality inventory. A Romanian indigenous measure based on the five-factor personality model

Sava, Florin A. (West University of Romania, Romania)*
Maricutoiu, Laurentiu P. (West University of Timisoara, Romania)

Abstract

DECAS (Sava et al., 2008) is a 95 item test that solicits dichotomic ("true" / "false") answers to items equally distributed among five content scales: openness, extraversion, conscientiousness, agreeableness, and emotional stability (as opposed to neuroticism). The aim of this oral presentation is to introduce briefly the main features and psychometric properties of this newly developed instrument to assess personality according to the well-known five-factor model. Data concerning norms and internal consistency have been collected from more than 15,000 participants, aged from 16 to 65 years. The most important validation study using gold standard instruments such as Costa and McCrae's NEO-FFI (1992) and Goldberg's IPIP (1999) supports the construct validity of DECAS. Data collected from other validation studies are also briefly presented, in particular those involving the criterion validity of DECAS scales. The results support good criterion validity for the various DECAS scales in relation to both real-life criteria (academic and job performance, penalties applied to automobile drivers for infractions, tendency to change employment) and laboratory criteria (persistence with tasks, punctuality at meetings).

C19-5

Indigenizing personality assessment: Possible, inevitable, or even desirable?

Hill, Jill (Teachers College, Columbia University, USA)*
Scharron del Rio, Maria (Brooklyn College, City University of New York, USA)
Skolnik, Avy (Teachers College, Columbia University, USA)

Abstract

Within the literature, there is consensus on the need for greater understanding of the inherent conflicts between Western and Indigenous concepts of health and pathology (Allen & Dana, 2004; APA, 2003; LaFromboise et al., 1991). A major part of developing such an understanding is to critically examine the assumptions that form the foundation of Western psychology, especially with regard to assessment. For many Indigenous peoples, the only purpose of Western-based assessment instruments, after more than 500 years of colonization, is to assimilate what's left – the psyche (Mohawk, 2004). More recently, in response to widespread cross-cultural adaptation and use of Western measures, researchers have developed Indigenous assessment measures, specifically personality assessment instruments, that have been shown to effectively measure personality attributes that are not considered by Western measures but valued as important Indigenous personality constructs within specific cultural contexts (Cheung et al., 2003). This presentation examines the feasibility and desirability of developing Indigenous measures of personality within the U.S. Indigenous Peoples in the U.S., comprised of more than 560 different Nations with more than 260 languages still spoken, are hardly a homogenous group. This is but one challenge to the development of Indigenous measures of personality. Other considerations and challenges include the historical mistrust of Western-based psychological interventions that have perpetuated imperialistic agendas within Indigenous communities. The presenters address other challenges to developing Indigenous measures of personality and explore concerns about further Western cultural proselytization (Gone, 2008) via assessment methods among Indigenous communities.

C19-6

'The princess diana' a new way to establish 'basic needs'

van Luijk, Frank (Free University, Amsterdam/LTP, Netherland)*

Abstract

A debate has reigned for many years concerning the way in which 'basic needs' should be diagnosed. McClelland, the great advocate of these concepts, said that the only proper way to do this was by using projective techniques. Others advocated the use of personality questionnaires. When the results of personality test and projective testing did not converge, those adhering to projective tests said this was a logical consequence of the use of the wrong technique, those who were in favour of the personality questionnaires said the same. 'The Princess Diana' is an alternative way of establishing basic needs. It is an interactive test in which the candidate is the captain of a large cruise ship. The candidate is confronted with different management problems, in each situation there are between two and six different possible reactions. Each reaction 'satisfies' a different need. After choosing one of the alternatives, there is a new situation, which is a logical consequence of the chosen alternative. This means that every candidate has his own unique test; the test is like a scenario, the exact story depends on the chosen alternatives. This means that for each candidate the length and the content of the test are different. Because of this 'classical' scoring is not possible. Information-theory was helpful in solving some of these problems. In this presentation examples will be shown of the different situations and alternatives and of the way the scoring is done.

C20 10:00-11:30 Room211

C20-1

Standardizing rater performance: Empirical support for regulating language proficiency test scoring

Ackermann, Kirsten (Pearson Language Tests, United Kingdom)*
Kennedy, Lauren (Second Language Testing, Inc., United Kingdom)
De jong, John (Pearson Language Tests, United Kingdom)
Zheng, Ying (Pearson Language Tests, United Kingdom)

Abstract

This paper is concerned with the standardization process during rater training and explores the need for a tight framework of empirically established rules to manage human scoring processes. This research employs data from surveys, questionnaires and face-to-face interviews carried out with the trainee raters and supervisors in the course of the field tests of the Pearson Test of English Academic in 2007 and 2008. The purpose of this paper was to analyse and evaluate the process of standardizing rater performance, share the findings, and make recommendations to further improve rater training. The following aspects were analyzed to determine the requirements and develop the means of standardizing human rating: the role of supervisors, recruitment and role of raters, organization of training, training materials, and management support. The results of the analyses and the evaluation of the standardization highlight the importance of close monitoring of human rating, given the need to avoid jeopardizing the validity of the associated test. The findings have implications for test development in general and the human scoring process in particular.

C20-2

Assessment of adult literacy: A French national survey

Megherbi, Hakima (University Paris 13, UTRPP, Villetaneuse, France)*
Rocher, Thierry (University Paris Ouest Nanterre La Défense, France)

Abstract

The French national survey 'Information et Vie Quotidienne' ("Information in Everyday Life") aiming to measure literacy performance in everyday life has been conducted in 2004 with about 10,000 adults aged from 18-to-65 year-olds. This survey was organized by the French Institute of Statistics (INSEE) in collaboration with the Ministry of Education (DEPP). It takes into account the critics made to IALS (International Adult Literacy Survey), especially by ensuring a more natural relation with interviewer to encourage motivation. With regard to the test, several documents have been elaborated based on the cognitive theoretical approach (Gernsbacher, 1990; Kintsch, 1998; Van Dijk & Kintsch, 1983). In this framework, reading comprehension is defined as a complex activity that consists of mentally elaborating a coherent representation. Several components that have identified in reading comprehension were addressed: word knowledge, syntactic and semantic processes of sentence, and text-integration. A total of five source documents varying in their structure type (Kintsch, 1998) have been elaborated: narrative, scientific, expositive texts, maps describing roots and bar charts documents. Respondents (7389 adults) had to read each source document and to answer some questions which were multiple-choice questions or open-ended questions. Multiple dimensional analyses revealed a bi-dimensional factorial structure: The items of the narrative, scientific and expositive texts loaded on a same factor and the items of the maps and bar charts documents loaded on another factor. To sum up, this survey reveals two specialized abilities (verbal and visuo-spatial) for the comprehension of text documents.

C20-3

Developing diagnostic tests for bilingual education: A case study of Irish immersion

Mulhern, Gerry (Queen's University Belfast, Northern Ireland)*
Wylie, Judith (Queen's University Belfast, Northern Ireland)

Abstract

Immersion education in indigenous minority languages, such as Welsh, Scots-Gaelic, Irish, Basque and Catalan has seen a significant growth in recent decades. For example, in Northern Ireland and the Republic of Ireland, approximately 38,000 children are enrolled in immersion schools outside the official Irish-speaking regions. Internationally, this development has given rise to issues relating to educational and cognitive assessment within the immersion/bilingual education sectors. When learning difficulties in particular are considered, it is necessary to distinguish between developmental characteristics associated with bilingualism and those characteristics associated with atypical development. Currently we are developing a series of diagnostic tests for use within the English-Irish immersion education sector. In this paper we present details of these tests and discuss issues relevant to psychometric testing within a bilingual immersion education context. We describe a recently versioned (Irish) and standardised working memory test battery. Patterns of first (L1) and second (L2) language performance in children aged 7-12 years, and issues concerning the discrepancy between L1 and L2 performance, will be presented. We also describe our ongoing project to develop diagnostic tests for executive functions, literacy and numeracy. Several issues identified in these development projects are discussed, including the need for versioning rather than translating of existing materials, and the importance of assessing children's cognitive abilities in each of their languages. These issues are particularly important in the context of assessing the learning difficulties of bilingual children, especially within an immersion education context.

C20-4

Comfort levels in using the English language as an indicator of English language ability across seven Asian regions

Lee, Tony (Assessment Research Centre, Hong Kong Institute of Education, Hong Kong SAR, China)*

Abstract

This paper aims to present a way to meaningfully quantify comfort levels in using the English language by foreign / second language users of English in their daily lives and the relationship between comfort levels and English language ability. The measure of comfort levels described in this paper was developed in conjunction with an extensive survey on language use patterns carried out in April 2005 by a large English language teaching network, the Wall Street Institute (WSI) in Hong Kong. The survey was administered to about 15,000 respondents in seven Asian regions: Hong Kong, Japan, Mainland China, Singapore, South Korea, Taiwan and Thailand. The instrument is a nine-item questionnaire forming part of a survey instrument with 43 items on various aspects of English language use in real life, e.g., use of English in work and leisure and dating. The comfort level items cover three English language use situations: communicating with native speakers of English, using English language media, e.g., TV, and reading English language materials. There is also one item in the survey instrument for respondents to self-assess their own English language ability. Using the Rasch measurement models, a system to index comfort levels in using the English language. The indexing system was sufficiently consistent and correlated highly with self-assessed English language ability. The indexing system of comfort levels in using the English language developed can be conveniently used as a quick-and-dirty measure of English language ability for a larger sample.

C20-5

Developing and evaluating instructionally sensitive assessments – why we need them and how can we develop them

Ruiz-Primo, Maria Araceli (University of Colorado Denver, USA)*

Abstract

Instructionally sensitive assessments are intended to provide a better picture of the extent to which instruction impacts student achievement and learning. Why we need these assessments and how we can develop them is the focus of this paper. More specifically, the paper reports on a pilot study with nine teachers in which an approach for developing instructionally sensitive assessments is being tested. The approach is based on variations in the proximity of assessments to the intended curriculum (close, proximal, and distal assessments; see Ruiz-Primo et al., 2002) and different facets of achievement (i.e., declarative, procedural, and schematic knowledge). The study reported intends to: (1) empirically test the approach for developing instructionally sensitive assessments; (2) indentify the critical components of instructionally sensitive assessments; and (3) provide empirical evidence about the robustness of the approach. The pilot study was carried out in three, school districts of different locale (urban and suburban) and used different fifth-grade science curricula. Teachers mapped their curricula and participated in a field study; science and assessment coordinators developed items, and an expert panel reviewed the assessment items. Items were administered in a pre-post-test design. The paper will provide information on the quality of the instructionally sensitive items developed based on the expert panel review, the psychometric characteristics of the items, and other empirical evidence on the robustness of the approach and the quality of the items.

C20-6

Creating a standard Arabic assessment

Thomas Ahluwalia, Nancy (Prometric, USA)*
Chelli, Karim (Dubai Education, Dubai)

Abstract

Students who attend institutions where instruction is provided in Arabic have sometimes lacked the necessary language skills to benefit from this instruction. Dubai Education has embarked on the development of a testing program to address the issue. The purpose of the testing program is to measure a candidate's level of language skill in reading, writing, and eventually listening. Prometric and Dubai Education have collaborated on the development of the testing program. The initial stage of the process was the development of the blueprint for the examination. A panel of experts from the United Arab Emirates, Saudi Arabia, Kuwait, and France participated in a session in which the test content was determined. The test will contain items to measure reading skills including skimming and scanning, writing and language skills, and listening. The introduction of these elements will be staged. Using paper and pencil delivery the reading and writing items were pilot tested with student groups in the United Arab Emirates. The listening component will be added later. The session will describe the challenges associated with developing a standard for Arabic and the preliminary steps taken to address these challenges.

Poster Sessions

Monday, 19th July 2010

P01 14:00-15:30 Room105

P01-1

The development of a computerized statistical literacy assessment for college students

Chao, Yu-Ning (National University of Tainan, Taiwan)*
Tzou, Hue-Ying (National University of Tainan, Taiwan)
Ding, Ching-Huei DING (National University of Tainan, Taiwan)

Abstract

Statistical concepts can improve our lives personally and professionally. College students should be prepared with statistical literacy to make reasoned judgments and form balanced opinions in the future. The purpose of this research is to develop a computerized statistical literacy assessment to evaluate the college students' statistical literacy performance. The researcher develops the measuring tool for statistical literacy which is based on Gal's statistical literacy model, and Watson's statistical literacy level theory. Besides the online news, market survey, and newspaper, multimedia like TV advertisement and news, is also used to develop the questions which in this tool are mainly related to daily life. Each question with multiple choice or open-ended question in this measuring tool is divided into four parts which include "mathematical / statistical knowledge", "context knowledge", "inference skill", and "critical knowledge". The coding of data according to Watson's statistical literacy level theory and adult literacy, and the Rasch model will be used to scale this tool. This research will provide a valid measuring tool to evaluate the college students' statistical literacy performance, and the findings are expected to be indications for following studies.

P01-2

A short version of Supports Intensity Scale (SIS): The utility of the application of artificial adaptive systems.

Gomiero, Tiziano (ANFFAS Trentino Onlus, Italy)*
Croce, Luigi (Catholic University (Brescia, Italy)
Grossi, Enzo (Medical Department Bracco SpA, Italy)
De Vreese, Luc Pieter (Psychogeriatric Service, Health District, Modena, Italy)

Abstract

The aim of this paper is to present a shortened version of the Support Intensity Scale (SIS) obtained by the application of mathematical models and instruments, adopting special algorithms based on the most recent developments in Artificial Adaptive Systems. All the variables of SIS applied to 1052 subjects with Intellectual Disabilities (ID) involved in the validation of the Italian version of SIS, were analyzed with the aforementioned Artificial Adaptive Systems. This study has identified 56 items, whose responses are able to explain up to 89% of sample variance. Secondly, these same variables have been analyzed by means of specific semantic networks, in order to demonstrate the plenty of scientific suggestions emerging from such an approach to statistic processing.

P01-3

Effects of different conditions for the translation of a test on the PC

Kreuzpointne, Ludwig (University of Regensburg, Regensburg)*

Abstract

The role of item presentation, dispatching the answer, and presentation of the processing-time in the context of equivalence of paper-based-testing (PBT) and computer-based-testing (CBT) was analyzed. 298 pupils (M=18.9, SD=2.5 years) were tested two times, one group first with the PBT and several weeks later with the CBT, another group the other way round. The used test battery (revised LPS; originally Horn, 1983) contains 11 subtests measuring crystallized and fluid intelligence, visual perception and cognitive speediness (three-stratum-theory; Carroll, 1993). The results subsumed for all subtests show, that it doesn't matter, if the items of the CBT were presented one after another or all together, if the kind of test is unknown by the subjects. If the CBT is the second one, the presentation of all items together will lead to a higher degree of equivalence. If there is a button to accept the marked solution or not as well as the opportunity of changing already given answers plays a minor part for the equivalence. It although plays a tangential role if the remaining processing time is shown or not. For the single subtests the influence of the conditions were partly different, especially when the rate of speed component is high. The higher the mean item-difficulty the smaller the influence of the variation of the condition. But the equivalence of PBT and CBT couldn't ensured with maximum similarity of both modes as well as it is possible that probably different seeming translations could lead to equivalent tests.

P01-4

Construct a vertical scale for Science Assessment

Kuo, I-Ting (Graduate Institute of Educational Measurement and Statistics, National University of Tainan, Taiwan)*
Twu, Bor-Yaun (Graduate Institute of Educational Measurement and Statistics, National University of Tainan, Taiwan)

Abstract

Scaling allows one to compare scores from different test forms, and vertical scaling is intended to support the comparison of scores obtained at each of a number of test forms (or levels) of systematically different difficulty. The primary reason for creating vertical scales is to measure learning across time. Many assessments, such as ITBS (Iowa Tests of Basic Skills), CAT (California Achievement Tests), SAT (Stanford Achievement Test), and so on, have created vertical scales in different ways, yet all of those scales appear to be functioning adequately for some of the same purposes. Scaling variations included growth definition, data collection design, scaling method, item response theory (IRT) scoring procedure, proficiency estimation method, and evaluating resulting vertical scales. The data for this study were extracted from responses to the 2006 Gifted Students Screening Assessment (GISA) science assessments in grade 3 through 6.The assessments were constructed with common item data collection design between adjacent grades to support the establishment

of a vertical scale, and then use IRT separate estimation to independently estimate the item parameters for each level being on a separate scale. Finally, average grade-to-grade growth, grade-to-grade variability, and separation of grade distributions have been used to evaluate the results of vertical scaling in this study.

P01-5

Providing validity evidence for fair use in international testing: A multi-group confirmatory factor analysis approach

Li, Ying (University of Maryland, USA)*
Jiao, Hong (University of Maryland, USA)
Lissitz, Robert (University of Maryland, USA)

Abstract

Background The International Testing Commission (ITC) developed ITC International Guidelines for Test Use (ITC, 2000) and required test users to give due consideration to issues of fairness in testing. It suggested providing validity evidence to support the intended use of the test in the various groups, including (1) demographic groups (e.g., gender, cultural background, or ethnicity groups), (2) language groups within or across countries, and (3) regular and disabled groups. Objective The objective of this study was to use the Multi-group Confirmatory Factor Analysis (MG-CFA) method to provide validity evidence for construct invariance across groups, so that test fairness issues can be addressed quantitatively and documented. Method MG-CFA was conducted to examine the subscore structure of a Maryland statewide biology test and the consistency of the subscore structure across non-accommodated (regular) and accommodated (disabled) students. Four hierarchical tests of construct invariance were conducted, from least restrictive to most restrictive: (1) configural invariance, (2) week invariance, (3) strong invariance, and (4) strict invariance. Chi-squared likelihood ratio tests and model fit indices (e.g. CFI, TLI, RMSEA, and SRMR) were obtained to determine the level of construct invariance. Results and conclusions The model fit indices indicated good fits for all four hierarchical tests of construct invariance across groups, and the chi-square likelihood ratio tests suggested that strict invariance was achieved. These provided evidence of construct invariance across groups. Implications The MG-CFA method of providing validity evidence can be widely implemented to any manifest group to ensure fair use in international testing.

P01-6

Development and application of online assessment for experimental debugging performance

Hung, Pi-Hsia (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)
Lin, Chien-Yu (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)
Lu, Cheng-Chin (Department of Education, National University of Tainan, Taiwan)*
Chang, Yung-Yin (Department of Education, National University of Tainan, Taiwan)

Abstract

This study looked at the development and application of online assessment for scientific literacy. The design focused on experimental error detection with sixth grade students as the research participants. A main goal of science education is to improve scientific literacy, an important part of which is power of observation. The experimental debugging ability is the important factor in the process to find questions of science education. An online assessment was designed for this study to develop a computerized assessment for the subject of natural science. Experimental debugging ability was tested through the observation of experimental steps in order to improve scientific literacy. The units of basic knowledge of experimental process and cell division were used to create sub-tests, The contents of the online assessment were presented mainly through text, images, animations and multimedia videos. The research results showed that testing in a multimedia format provided an advantage in terms of scenario perception.

P01-7

Internet platform TOP-diagnostic: Experience of cross-cultural adaptation and using for counseling

Ritz-Schulte, Gudula (Osnabruck University, Germany)
Kuhl, Julius (Osnabruck University, Germany)
Mitina, Olga (Moscow State University, Moscow City University Psychology and Education, Russia)*

Abstract

The concept is based on PSI theory (personality systems interaction theory), an integrative attempt to unify advances in cognitive, motivational and neuropsychology. PSI theory permits a deep understanding of clients' intrapersonal functional dynamics, especially need management and self regulatory competences. The special diagnostic system TOP (Training oriented Personality Assessment) is used through internet access, provides a micro-analysis of cognitive, emotional, motivational and self-regulatory functions. It allows to conduct personality oriented counseling, takes into account many significant psychological personality variables of a client, in a functional analytical way. POC substantially increases counseling efficiency because it is based on a comprehensive micro-analysis of the client`s personality functioning. The main TOP modules allow to measure and to study personal styles, self- management, self-regulation, explicit and implicit motivational styles, explicit- and implicit mood and mental states, mental and psychological symptoms. As interpretation the psychotherapist gets the conclusion and many visual graphs which allow to represent results and client's problems easy and understandably. Target groups are managers, sales managers, team leaders, assessment center, students, pupils, patients. The TOP originally created in German now is translated into Russian and English. The experience of adaptation and using TOP in counseling will be presented along with cross-cultural features revealed during the work process. Several research projects in Germany and Russia that validate TOP assessment and the way of counseling derived from it will be presented.

P01-8

Current use of psychometric tests for selection in China – results of the 2010 China Talent Selection Survey

Morley-Kirk, James (China Select, China)*

Yang, Louis (China Select, China)

Abstract

This paper presents a picture of the current use of psychometric tests in assessment programmes to choose between candidates applying for employment vacancies in China. Data will be drawn from the results of the 2010 China Talent Selection Survey and illustrate main practices and attitudes towards psychometric tests overall, across industry sectors, and across organisation sizes. Comparison with data from 2008 and 2009 surveys will suggest trends. Implications for psychometric education and for test publishers will be considered.

P01-9

Development and application of multimedia intergrated assessment for sense of pitch and rhythm

Hung, Pi-Hsia (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)

Lin, Chien-Yu (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)

Ni, Chen-Hao (Department of Education, National University of Tainan, Taiwan)*

Chou, Yu-Yin (Department of Education, National University of Tainan, Taiwan)

Abstract

The purpose of this study is to explore the intergration of online assessment applications in multimedia. With sixth grade students as the test participants, computerized online assessment for sense of pitch and rhythm were developed to assess the learning effectiveness on music subject. This study of musical online assessment focused mainly on the sense of pitch and rhythm. The online assessment was designed to make use of images, animations, videos and music with software programming used to provide an interactive multimedia interface that is distinct from conventional musical assessments. There are there states of interactive multimedia function: cognition, learning and perception. Pitch and Rhythm were used to create sub-tests, This study is devoted to discover the distinct characteristics of interactive multimedia. Based on the results of the study, it was concluded that the intergration of multimedia in online assessment made the learning process more interesting and improved the applicability and future development.

P01-11

The impact of instant feedback during unsupervised online aptitude testing

Preuss, Achim (cut-e GmbH, Germany)*

Wehrmaker, Maike (cut-e GmbH, Germany)

Abstract

New technology allows completely new methods and structures of cognitive ability tests. Unsupervised aptitude testing requires extended quality aspects that go beyond classical psychometric criteria. In particular the instruction sequence has to be constructed in a way that takes into account individual prerequisites of participants to ensure fair testing. However, beliefs of self-efficacy have a significant impact on performance in aptitude testing. Self-efficacy beliefs are not only influenced by the feedback to example tasks in the instruction sequence but also during a test by the perceived difficulties of tasks and the individual assumptions about correct and incorrect responses. In order to analyse the impact of instant feedback during unsupervised online aptitude testing participants of a broader validation study (N=602) were randomly allocated to different feedback versions of an online logical reasoning test. The results indicate a significant impact of feedback variants on performance and the reliability of the test itself.

P01-12

The effectiveness of Computerized Assessment Technique (CAT) versus paper and pencil in the assessment of executive functions for brain injury.

Roberts, Patricia (University of Bedfordshire, UK)*

Ertubey, Candan (University of Bedfordshire, UK)

Teoh, Kevin (University of Bedfordshire, UK)

Abstract

This research was informed by the notion of executive functions included in Baddeley's (1986; 2007) working memory model. This model includes subsystems specialized for the maintenance of speech based information, the phonological loop, for visual and spatial information the visuospatial sketchpad, and a central control, the central executive. The aim was to use computerized tests and their paper pencil equivalents to establish if electronic versions are a competitive alternative for assessment of executive functions in an adult population. Using an experimental method, differences in performance from both the presentation (Paper-Pencil versus CAT) and groups (control versus patients with acquired head injury) was investigated. 27 participants took part in this study (M = 34.04, SD = 13.68). CANTABeclipseTM, a recent development, which facilitates the computerized assessment of cognitive deficits as a result of, frontal lobe damage, stroke, and neurodegenerative diseases. CANTAB is made up of 22 subtests in 6 domains. This research investigated the CANTAB test battery (5/22 subtests) versus paper-pencil test equivalents (i.e. Tower Honai, WCST, WAIS) on a normal population and patients suffering from frontal lobe damage. The results from the control and acquired head injury group were compared for each of the CANTAB subtests and paper-pencil equivalents. Overall 3/5 CANTAB subtests and 4/5 paper-pencil tests showed significant differences between groups. No interaction effect was found. The results are seen as promising for the use of CAT in the identification of frontal lobe deficits. The advantages and disadvantages of the both electronic and paper-pencil tests are discussed.

P01-13

Psychometrics and paradigmatic shifts: Ethical and epistemological considerations

Scharron-del Rio, Maria (Brooklyn College - City University of New York, USA)
Hill, Jill (Teachers College, Columbia University, USA)*

Abstract

Instruments are developed, whether explicitly or implicitly, within a theoretical framework, supported by multiple assumptions, and constructed within a particular worldview and historical time. This worldview and context contain particular political structures that privilege particular assumptions regarding what constitutes an appropriate object of study, how it is defined, and how it should be studied. According to Potts & Brown (2005), it is imperative to make these epistemological political practices explicit in order to engage in anti-oppressive and empowering research. The Western "discourse of discovery" (Linda Smith, 1999), assumes that knowledge is "created" or "discovered" objectively, regardless of the worldview of the researcher or its context. This approach results in a decontextualized, compartmentalized, and partialized account of the human experience, as the complex interactions between context and object of study are minimized or ignored (Barnhardt & Kawagley, 2005; Martín-Baró, 1998; Smith, 1999). This worldview is often at odds with how diverse ethnocultural communities see themselves and their world. How do we avoid cultural imperialism within the realm of research; What would be an ethical and multiculturally competent approach to the assessment of ethnocultural communities whose worldview is not reflected by the instruments traditionally used to assess them; How can psychometric research contribute to the self-determination of disenfranchised ethnocultural and indigenous communities? The presenters (one Puerto Rican and one American Indian) will address these questions as well as the challenges related to navigating competing knowledge systems within the confines of their careers within Western psychology and higher educational institutions.

P01-14

Job applicant screening with adaptive classification tests of personality

Gnambs, Timo (University of Linz, Austria)*

Abstract

International companies implementing web-based recruiting procedures increasingly rely on screening procedures to select those applicants for further consideration (e.g. for face-to-face interviews) who exhibit a minimal level of an elemental trait ("negative screening"). In these cases an individuals ´ accurate standing on the latent trait is of only secondary concern as long the test achieves accurate classifications of the applicants into those who fall below or above a certain predefined threshold. To date, adaptive classification testing has been dominated by research on dichotomous response formats and classifications into two groups (for pass/fail decisions). This paper extends this line of research to polytomous classification tests for two and three group (e.g. inferior, mediocre and superior proficiency) scenarios. It is demonstrated that even for established personality scales with a polytomous response format, the average test length can be significantly reduced by applying an adaptive classification procedure. Two computer simulations with generated (N=10000) and real responses (N=2000) to established personality scales (conscientiousness, achievement motivation, opinion leadership) of different length (12, 20 or 29 items) confirm that adaptive item presentations significantly reduce the number of items required to make such classification decisions by a third to a half, while maintaining a consistent classification accuracy. Additionally, a simulation experiment demonstrates that (a) the choice of the stopping rule and (b) the number of classification groups have a significant impact on the average test length.

P01-15

Just because the tests don't match

Tono, Suwartono(Muhammadiyah University of Purwokerto, Indonesia)*

Abstract

A good test employs theoretical and practical considerations in its construction. Unfortunately, in relation with this, many tests, including large-scale ones, fail. In Indonesia, the ability to speak and write in English is the most obviously demanded. At least, our advertised vacancies commonly require that job seekers have proficiency in these two language skills. It is not surprising, because in this modern era, these two language skills play a vital role in many fields of life. However, our prominent tests of English do not measure these skills proportionally. Another serious matter is that the majority of our tests use just one test technique, i.e. multiple choice, for the whole test. The risk of bias increases with less varied techniques. Many more of the current testing practices are well worth criticizing. This paper dedicates itself to this. It starts with a brief theoretical review on test backwash and logic validity. Then, it presents the shortcomings existing in the National Tests of English, along with the consequences that might arise, including unsatisfactory skills learning outcome. And, it ends with recommendations addressed to relevant authorities.

P01-17

Online assessment of virtual reality is integrated into wayfinding performance

Hung, Pi-Hsia (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)*
Lin, Chien-Yu (Graduate Institute of Assistive Technology, National University of Tainan, Taiwan)
Tu, Jia-Ling (Department of Education, National University of Tainan, Taiwan)
Hong, Pei-Tsen (Department of Education, National University of Tainan, Taiwan)

Abstract

Wayfinding is an important indicator when testing spatial ability and the use of 3D virtual reality allows online assessment to be of special value and contribution. This study is the development of an online assessment of virtual reality is integrated into wayfinding performance. Decision-making for wayfinding problems needs relative information about the space. The multimedia design includes a combination of text, image, 3d models, animation, video. In this study, a wayfinding video in 3D virtual reality was used to design an assessment. Landmarks and the presentation of the video from different angles were used to create two subtests. All the participants of this study were 6th grade students. The purpose of this study is to investigate users' wayfinding behaviors and spatial knowledge in virtual environment. The results of the study showed that wayfinding performance varied according to the spatial information conditions. Landmarks helped users establish better wayfinding performance in a virtual environment.

P01-20

Making classification decisions in cognitive diagnostic assessment using a new estimating function of individual's attribute mastery probabilities

Zhang, Shumei (School of Mathematical Sciences,Beijing Normal University, China)
Sun, Jianan (School of Mathematical Sciences,Beijing Normal University, China)*

Abstract

Rule Space Method (RSM) can make classification decisions by individual's knowledge state and obtain individual's cognitive diagnosis result. Although it mentioned how to estimate individual's attribute mastery probabilities, the method didn't use those probabilities to make classification decisions. Firstly, we explore a new cognitive diagnostic method which uses a model we developed to estimate individual's attribute mastery probabilities, and then makes classification decisions using cluster analysis with those attribute mastery probabilities. A simulation study was employed to compare this cognitive diagnostic method with RSM. Finally, real data analysis for comparing the two methods was presented. Specifically, based on item response pattern and Q matrix, we built a new estimating function of individual's attribute mastery probabilities, and use an iterative algorithm to acquire the estimates of those probabilities.

P01-21

Formation and control of neutralization in subjective rating

Wang, Bo (School of Psychology, Beijing Normal University, China)*
Che, Hongsheng (School of Psychology, Beijing Normal University, China)
Song, Ke (School of Psychology, Beijing Normal University, China)

Abstract

Subjective items are widely used in many large-scale personnel selections and educational tests. The recently broadly applied online scoring system makes it convenient to control the errors in subjective rating. However, through the analysis of the rating process and results of one large-scale personnel selection, it is found that in order to reach the consistency, raters tend to be conservative about the rating and there appears obvious neutralization in the score distribution. The neutralization is caused both directly and indirectly by the rating standards, the cognitive patterns while rating and the error discrepancy setting between raters. Longford's method and multi-faceted Rasch model are both applied to adjust the scores and rating augmentation pattern is also used to control the rating process. Both ways have reduced the neutralization and improved the rating accuracy without impacting the rating consistency. Key words: subjective rating, neutralization, Longford's method, multi-faceted Rasch model, rating augmentation pattern.

P02 15:45-17:15 Room105

P02-1

A validity study on the cartoon integrated computerized English listening test for Taiwanese students

Chen, Su-Yu (National University of Tainan, Taiwan)*
Hung, Pi-Hsia (National University of Tainan, Taiwan)

Abstract

English is the major language of business, technology, science, the Internet, popular entertainment, and even sports (Nunan, 2003). Promoting every citizen's English proficiency to ensure Taiwan's competitiveness in the global arena is a very important policy. English has become a required subject in 3rd grade since 2005. The rapid spread and infusion of technologies such as computers, digital media, the Internet, video and audio recorders, and playback devices into every phase of contemporary life is affecting both the methods of learning and assessment and the content of what needs to be learned in schools. The purpose of this study is to develop a cartoon integrated computerized English listening test (CICELT). The innovative item format developed allows more operant responses. The item types included in CICELT are film, flash and static pictures. The correlation coefficients between three item types and students' English grades of school and standardized reading comprehension scores are discussed as the preliminary construct validity evidences. The characteristics of students of ASAP (after school alternative program) on CICELT are described. The implications for intervention design are also included.

P02-2

A critical issue for cognitive diagnostic test

Ding, Shu-Liang (Jiangxi Normal University, China)*
Wang, Wen-Yi (Jiangxi Normal University, China)
Yang, Shu-Qin (Fujiang Normal University, China)

Abstract

How to establish the relationship between the observable response pattern (ORP) and the latent knowledge state (LKS) is a critical issue for cognitive diagnostic test (CDT). Tatsuoka (1991,1995,2009) employed Graph Theory, Boolean Algebra and other tools to develop a Q matrix theory. It is easy to make a counterexample to her conclusion about the Boolean lattice (BL). For example, in the Example 4 (Tatsuoka, 1995,p.333) or in the Example 4.1 (Tatsuoka,2009,pp.88- 89), the distributive law is not satisfied, and the BL could not be constructed. In the Example 3.5 (Tatsuoka,2009, pp.74-75), there are only 6 different nodes, so it is not a BL neither. There is a scheme to deal with the problem how to mend Tatsuoka's Q matrix theory. Suppose that the attributes and their hierarchy are given, the reachability matrix could be calculated. Then the matrix Qs which is a set of all of the possible LKS could be derived from the reachability matrix using the augment algorithm (Ding et al.2008,2009;Yang et al.2008). This blueprint is represented by a matrix, Qt, whose columns are drawn from the matrix Qs, all ideal response patterns (IRPs) except zero vector could be calculated based on Qs and Qt by straightforward algebra computation. So the relationship of ORP and LKS is obtained. Some other important applications could be drawn from the augment algorithm,such as test construction,test equating, validity for CDT and strategy of selection item for computerized adaptive testing with CDT.

P02-3

IRT equating for multidimensional data

Xie, Jing (Education College, Capital Normal University, China)
Fang, Ping (Education College, Capital Normal University, China)*
Wang, Peng (Education College, Capital Normal University, China)
Ma, Ying (Education College, Capital Normal University, China)

Abstract

IRT equating method must satisfy the hypothesis of unidimension, but the real test usually is multidimensional that match the experience that completing a examination always needs many kinds of abilities coordinate together. This research is to bring matured IRT equating method into inspecting multidimensional data, examine its stability, and try to spread its application scope. Five study involved in this research. The main conclusions were as follows:(1) When equating multidimensional data, IRT equating method is still stable.(2)The equating results of 4 IRT transforming methods exist difference. The results of mean/sigma and TTC are more accurate and stable, but mean/mean methods is not suitable to process multidimensional data.(3) The equating results of 10 population exist difference, more disperse, more worse. (4) The equating results of 12 anchor exist difference. Among them, the effects of item construct and type are more prominent, but anchor item length is not significant.

P02-4

A survey on application of formative assessment criteria by teachers and relationship with students' academic achievement

Ghadimi Moghaddam, Malek Mohammad (Islamic Azad University of Torbat-e-Jam, Iran)*
Hoseini Tabatabaee, Foozieh (Islamic Azad University of Torbat-e-Jam, Iran)

Abstract

A survey on application of formative assessment criteria by teachers and relationship with students' academic achievement Abstract: The purpose of this research was to study the rate of awareness and application of criteria of formative assessment by teachers. The research method was descriptive survey tape. The subjects were the teachers who teaches in third grade of middle school in Razavi Khorasan's Province (15748 teachers and their students) and The sample was consists of 736 teachers (362 women and 374 men) that selected by stratified random sampling. They completed a questionnaire consisting of 24 questions that the coefficient alpha was .89. Data were analyzed by using: T- Test, correlation coefficient, ANOVA and Chi -square. Results showed that: There was difference between the teachers that participate in formative evaluation course were more awareness than others about criteria of evaluation. While 60.5% of teachers were unfamiliar with the principles of formative assessment, 39.5% were aware, out of which 27.6% did not apply the criteria of formative assessment in class. And only 11.9% of teachers under study used these criteria in classroom. There was significant relationship between teaching background, academic achievement and the rate of awareness criteria of formative evaluation. There was relationship between participation in formative evaluation course and application of criteria of formative evaluation.

P02-6

Hierarchical item response theory model with nonparametric prior distribution

Kuo, Bor-Chen (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)*
Hsieh, Tien-Yu (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)
Wu, Huey-Min (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)

Abstract

In this paper, a modified version of MH-within-Gibbs sampling method (Tierney, 1994) is proposed for estimating parameters of hierarchical item response theory (HIRT) model. This modified version is based on kernel smoothing which is the kind of nonparametric method. From simulation study, we find that the performance of MH-within-Gibbs sampling is poor when the distribution of incidental parameter is not normally distributed. In this paper, a simulation experiment based on HIRT model with MRCMLM is conducted to compare the performances of the parametric algorithm and the proposed nonparametric algorithm. In the experiment, three types of distributions of incidental parameters (normal, bi-mode and skewed distributions) are considered. The root mean square error (RMSE) of ability and item parameters is used to evaluate the performances of the parametric algorithm and the nonparametric algorithm. Experimental result shows that RMSEs of both ability and item parameters of the proposed algorithm are less than those of the parametric algorithm when the distribution of incidental parameter is bi-mode and skewed distributed. Besides, one real data is used for verifying the performance of the proposed version algorithm.

P02-7

Testing factorial invariance in multilevel data: A Monte Carlo study

Kwok, Oi-Man (Texas A&M University, USA)*
Kim, EunSook (Texas A&M University, USA)
Yoon, Myeongsun (Texas A&M University, USA)

Abstract

Testing factorial invariance has gained more attention in different social science disciplines recently. Nevertheless, when examining factorial invariance, it is generally assumed that the observations are independent from each other, which may not be always true. In this study, we have examined the impact of testing factorial invariance in multilevel data, especially when the dependency issue is not taken into account. We have considered a set of design factors, including: number of clusters, cluster size, intra-class correlation (ICC), and the magnitude of non-invariance at different level. The simulation results showed that the test of factorial invariance became more liberal (i.e., inflated type I error rate of testing invariance between groups) when the dependency was not considered in the analysis. In other words, it is more likely to conclude that factorial invariance does not hold when the factor structure is truly invariant across groups. Additionally, the magnitude of the inflation in the type I error rate was a function of both ICC and sample size (i.e., type I error rate became larger as ICC and sample size increased). Implications of the findings and limitations are discussed.

P02-8

Web-based and paper-based versions of the Reader Self-Perception Scale: A comparison of measure efficiency and participants perceptions

Leeson, Heidi (School of Teaching, Learning and Development, Faculty of Education, University of Auckland, New Zealand)*

Abstract

With the increased availability and affordability of computer processing power, graphical software, and high speed internet connections, the opportunities for presenting innovative items online is becoming a more viable proposition. However, the increased development required, both technical and psychometric, might prove a stalwart in the uptake of an online environment to administer such items. At the psychometric level a question that might be asked is how much measurement precision is provided by web-based items in comparison to their paper-and-pencil counterparts; Bluntly, is the extra effort and cost required to develop innovative items worth it? While literature has focused on establishing the validity or equivalence of measures administered under different modes, there exists little empirical psychometric information relating to the amount of information or reliability provided by innovative web items. In this study, the measurement precision provided by nine interactive scenario-based web items and their original paper-based versions was compared, with graphical and statistical fit analysis finding that the web items produced more measurement efficiency, particularly when items were shorter in length, and amongst students showing low reading self-perception. Thus, the wordier items required less perceived proficiency in reading to respond positively on the web, but were less discriminatory in this mode. At the test level, TIFs showed that innovative items were providing more measurement information across a wider theta range than the paper items, and functioning as though it was nearly 20% longer than the paper version.

P02-9

Nonclassical view on psychological assessment: New variables and new targets

Leontiev, Dmitry (Moscow State University, Russia)*

Abstract

The widely acknowledged targets of psychological testing are traits, states and abilities. These targets are based on the traditional view on human being as a relatively self-consistent creature whose behavior can be deduced from inborn causal mechanisms and external pressures and valences. Recent theoretical development shows that this view is at least incomplete. Besides traits, states and abilities, a number of more complicated targets emerge that refer to higher levels of human psychological organization. Among them are, e.g., (a) Contents = optional qualitative elements of the inner world (e.g. meanings, values, beliefs, attitudes) (b) Interpretations and attributions = optional contexts for experience processing (e.g. attributional style, locus of control) (c) Strategies of action, cognition, and experiencing = optional guiding directions for external and internal activity (action vs. state orientation, defenses and copings, hardiness, choice strategy) (d) Systems of varied structural complexity that arrange elements like motives, goals, meanings, personal constructs, or personal strivings (cf. Kelly, 1955; Emmons, 1999; Leontiev, 2007 etc.). These targets require a more complicated testing strategy than the traditional psychometric one, because their individual differences are hardly reducible to easily measurable scores; they are more of qualitative than of quantitative nature and hardly can be treated as normally distributed. Though there are various testing tools used for these kinds of targets, some of them rather well-known, their methodological reflection is still lacking. These targets provide new challenges and open new possibilities for testing methodology in research and assessment.

P02-11

Estimating the variability of estimated variance components for generalizability theory based on non-normal distribution data

Li, Guang-Ming (Center for Studies of Psychological Application, South China Normal University, China)
Zhang, Min-Qiang (Center for Studies of Psychological Application, South China Normal University, China)*

Abstract

Estimating variance component, which is the essential technique forgeneralizability theory, is constrained by sampling. Different sampling may cause different estimated variance component. Therefore, estimating the variability of estimated variance components needs to be further explored. In the past studies, there were some problems as follows. Fist, it was often that some researchers only focused on normal distribution data and neglected non-normal distribution data. In fact, non-normal distribution data could be always seen in such tests as TOEFL test. Second, the previous studies didn't compare the variability of estimated variance components using traditional, bootstrap, jackknife and Markov Chain Monte Carlo method (MCMC) at the same time. Finally, because there were not nearly researchers who studied more two non-normal distribution data at one time, there wasn't any research about interdistribution study for non-normal data. The study adopts Monte Carlo data simulation technique to compare the variability of estimated variance components for generalizability theory based on non-normal distribution data using four methods. The variability includes standard error and confidence interval. This result shows that only bootstrap method is good to estimate the variability for three non-normal distribution data and bootstrap method can do well between these non-distribution, but traditional, jackknife and MCMC cann't. This result also shows that bootstrap method need use "divide-and-conquer" strategy to obtain good estimated standard errors and estimated confidence interval and a uniform rules can be made as follows: boot-p for person, boot-pi for item and boot-i for person and item.

P02-12

Impact of outliers on decisions about the number of factors in exploratory factor analysis

Liu, Yan (University of British Columbia, Canada)*
Zumbo, Bruno D. (University of British Columbia, Canada)
Wu, Amery D. (University of British Columbia, Canada)

Abstract

Exploratory factor analysis (EFA) is a commonly used technique for identifying latent variables in assessment research. The decision about the number of factors underlying a set of observed variables is an essential part of EFA. Because of their availability in popular software programs (e.g., SPSS and SAS), two strategies have been widely used to help in determining the number of factors to extract: the Kaiser-Guttman (K-G) rule (i.e., eigenvalue-greater-than-one), and the Chi-squared test of fit (adequacy of the number of factors) that comes with maximum likelihood (ML) extraction in EFA. Factor analysis, like all statistics, may be influenced inappropriately by outlying data points. Outliers may have a large influence on the estimated models and their parameters. Some studies have investigated the effects of outliers on estimates in factor analysis and covariance models (e.g., Bollen & Arminger, 1991; Yuan & Bentler, 2001, 2007); however, there has been no systematic study of the effects of outliers on factor extraction methods in EFA. The purpose of the present study is to investigate how outliers affect decisions made by the K-G rule and Chi-squared test of ML using a Monte Carlo simulation. The simulation methodology described by Liu and Zumbo (2007) was used. Our preliminary findings indicate that outliers inflate the number of factors selected by the K-G rule and ML. Interestingly, in many situations, ML is more robust to outliers than the K-G rule. The present study aims to help researchers and test developers understand in which circumstances outliers are most problematic.

P02-14

Bayesian analysis of test reliability

Lozano, Luis M. (University of Granada, Spain)*
De La Fuente, Emilia I. (University of Granada, Spain)
Canadas, Guiliermo A. (University of Granada, Spain)
Vargas, Cristina (University of Granada, Spain)

Abstract

Frequentist procedures to estimate the reliability of a questionnaire are well developed. Lots of papers can be found were evaluating the properties of the different coefficients. One of the problems of the frequentist coefficients is that the researcher always starts from a point where no initial information is applied to the data. One of the advantage gained from using Bayesian techniques is the ability to incorporate disparate but relevant information into the analysis by way of the prior distribution. Another advantage is the ability to combine data from different analyses by using the posterior distribution obtained from a previous analysis as the prior distribution of the next analysis. These properties can be very useful for the test developers to evaluate in a more accurate way the reliability of the questionnaires. The aim of this work is to show how to estimate the reliability of a questionnaire from the Bayesian perspective using MCMC methods and compare the results with the ones are estimated from the frequentist framework. To reach this objective the SPSS and WinBugs software are used. This work has been founded with the project SEJ2006-13009 of the Ministry of education and Science and the project P07HUM-02529 of excellence of the Junta de Andalucía

P02-15

The comparison of scoring subscales approaches

Lu, Szucheng (National University of Tainan Graduate Institute of Measurement, Taiwan)*
Twu, Bortaun (National University of Tainan Graduate Institute of Measurement, Taiwan)
Wu, Cltunhsine (National University of Tainan Graduate Institute of Measurement, Taiwan)

Abstract

Many educational and psychological tests are designed with several subscales. Traditionally, researchers tended to estimate an overall ability (or overall score) using unidimensional item response theory to represent a examinee's performance and ignored subscales' ability. The practice has been changed recently. Several approaches had been propose to calculate the subscale score while also providing the overall score scale. The purpose of this study is to survey these methods and compare the results given by these methods with the traditional composite score obtained from raw score approach. Five methods for estimating item response theory scale scores for multiple subscales were surveyed. These methods included three multidimensional item response theory models: Bayesian MIRT model with a hierarchical structure (Huang, 2009), Rasch model with subdimensions (Brandt, 2008), and Bi-Factor Model (DeMars, 2006). Another method is a score augmentation approach (Wainer, Vevea, Camacho, Reeve, Rosa, Nelson, Swygert & THissen, 2001), which can be used with both classical test theory and IRT-based scores. The raw score method discussed in Wainer & Thissen (1993) will also be discussed.

P02-16

A cross-classification IRT model for differential item and testlet functioning

Wang, Aijun (University of Georgia, USA)*
Cohen, Allan (Univeristy of Gerogia, USA)

Abstract

Although item-level DIF analysis is the standard practice in educational testing, some researchers argued that differential testlet functioning should also be examined. However, fitting a standard IRT model to the testlet-based data violates the assumption of local item independence (LID), thus overestimating the preicison of ability and producing biased item parameters (Wainer & Wang, 2000). Therefore, a more complex model which accommodates the correlations among items within a testlet is needed. Noortgate, De Boeck and Meulders (2003) proposed a cross-classification multilevel logistic model which treats both items and persons as random. Under this model, examinee's responses are nested under pairs of item and person. Noortgate and De Boeck (2005) applied this model to DIF analysis. A group indicator is included in the model and the DIF effect is captured by the interaction between group and items. Stimulated by Noortgate and De Boeck's model, the current study applied this model to the testlet-based data to detect item-level DIF and testlet-level DIF simultaneously. Group indicator is included at the item and testlet levels and significant interactions indicate DIF effects. Three items and three testlets were found to function differentially. The testlets in which the DIF items were nested didn't show testlet DIF and the DIF testlets have no DIF items within them. The results indicate DIF items may not show testlet DIF when aggregated, and differential testlets may not have differential items within them.

P02-17

Comparison of classification accuracy and consistency between AHM & DINA

Wang, Wen-Yi (Jiangxi Normal University, China)*
Ding, Shu-Liang (Jiangxi Normal University, China)
Pan, Yi-Rao (Jiangxi Normal University, China)

Abstract

Diagnostic assessments, contrasted with summative testing which evaluates the student after instruction is over, are formative assessments. They are used to directly support teaching and learning. How to select appropriate psychometric model is a challenge to diagnostic assessments. The cognitive diagnostic models are divided into continuous latent trait IRT models, Bayesian networks models, IRT model-based sum-scoring, latent class models and Rule Space Approach (DiBello and Stout,2007). Classification methods are vital part of cognitive diagnostic models. The article mainly discusses classification accuracy and consistency based classification methods of AHM (a variation of Tatsuoka's Rule Space Approach, Leighton et.al,2004) and DINA (a latent class model, Junker et.al,2001).It also analyzes the leading cause of different results. The article gives a description why ideal response patterns with similar abilities are liable to make misclassification with classification method A of AHM (maybe appears in RSM). The simulate results , indicating the classification accuracy of DINA is higher than that of AHM (almost 10 percent) and high classification consistency between AHM and DINA , are supported by conclusion above. It also reports consistency of classification results from analyzing real data sets using AHM and DINA. These findings provide the reason for rational selecting AHM and DINA. Further studies show that traditional IRT methods could provide appropriate item parameter initial values for EM algorithm when estimating the parameters in DINA.

P02-18

How test design and uses influences test preparation: Testing a model of washback

Xie, Qin (Faculty of Education, Hong Kong University, Hong Kong SAR, China)*

Abstract

This study tested a model of washback. Washback on learning was conceptualized on the basis of the Expectancy-Value theory as the influence of test-takers' perceptions of assessment on their test preparation. Perceptions of test design and use, and prior language proficiency were used as predictors for test preparation; test expectation and values mediate between perceptions and test preparation. This study was conducted between April and June 2009 in the Mainland China, involving 1003 university undergraduates. Before that, a qualitative and a pilot study were conducted to develop and verify a measure of perceptions and a measure of test preparation practices. The two measures were given one at the beginning and one near the end of a two-month test preparation for College English Band 4. Prior language proficiency was also measured before the test preparation. Structural Equation Modeling was used to test this washback model statistically. This model turned out excellent model fit indexes supporting the hypothesized relationships: both perceived test design and test uses influenced test preparation; perceived test design influenced test preparation via the cognitive aspect of motivation-test expectation; perceived test uses influenced test preparation via the value aspect of motivation-test value. Implications for test designers, test users and test takers were drawn at the end.

P02-19

Comparison of Multivariate Generalizability Method and Guilford Method in estimating composite score reliability

Yang, Zhiming (Educational Testing Service, USA)*
Huang, Lanxiang (University of Texas at San Antonio, USA)

Abstract

Estimating composite score reliability is one of the routine tasks in test development, and work efficiency can be dramatically improved by using the appropriate estimation method. However, the current composite score reliability estimation method used in clinical assessment, the Guilford method (Guilford, 1954; Nunnally & Bernstern, 1994; Feldt & Brennan, 1989), is not very efficient because it prefers to have identical subtest standard deviations. This requires developing norms to produce grade/age norm-referenced scaled scores before estimating composite score reliability. However, developing clinical test norms is very time-consuming, and may also cause a lot of rework if the final composite score reliability turns out to be not psychometrically sound. This study investigates the similarities and differences between the Multivariate Generalizability Theory (mGT) method and the Guilford method in estimating composite score reliability. A clinical diagnostic reading test's composite score reliability is estimated by using both methods. Results indicate that the mGT method yields identical, or even slightly higher, composite score reliability estimates than the results from the Guilford method for the single-facet cross test design. This difference may be caused by the subtest scaling procedure, as mGT analysis is based on raw score and Guilford results are based on scaled scores. Unlike the Guilford method, mGT allows the composite score reliability to be estimated before the subtest norms are available. Therefore, the mGT method is more efficient than the Guilford method in psychological testing scale development practice.

P02-20

Evaluating the commonly used estimation methods in testing factorial invariance with ordinal measures methods in testing factorial invariant

Yoon, Myeongsun (Texas A&M University, USA)*

Abstract

Although factorial invariance now receives more attention in substantive research as well as in measurement literature than before, current studies have been mostly conducted with continuously measured variables. Factorial invariance with ordinal measures is less well-known, compared with continuous measure cases. In this paper, a simulation study was conducted to examine the Type I error and statistical power in testing factorial invariance with ordinal measures while comparing the three commonly used estimation methods in various simulation conditions. Additionally, the performance of the chi-square difference testing based on different estimation methods was evaluated in terms of finding non-invariant items. The results showed that Type I error rates were too high for the Robust Maximum Likelihood (RML) estimation method even with 5-category polytomous items, while Robust Weighted Least Squares (RWLS) yielded acceptably low Type I errors across all simulation conditions. Furthermore, power was adequately high only for polytomous item conditions but too low for dichotomous item conditions across all three estimation methods. Finally, the chi-square difference test using Weighted Least Squares (WLS) and RWLS performed well in detecting the non-invariant items while not over-identifying the invariant items as non-invariant for the polytomous item conditions. On the other hand, most of the non-invariant items were remained undetected for dichotomous item conditions when the chi-square difference test was used.

P02-21

What makes you stay? Money or affection: A comparative study of 19 cities in China

Zheng, Rui (Institute of Psychology, Chinese Academy of Sciences, China)*

Abstract

This paper addresses the debate concerning about the influence of money and affection on the urban identity. The results of the HLM analysis indicate that both money (GDP & Average income) and affection (Interpersonal Trust, Family Cohesion and so on) could work together.

P02-22

An extension study on Rule Space Model---From dichotomous to polytomous item

Tian, Wei (Institute of Developmental Psychology in Beijing Normal University, China)
Xin, Tao (Institute of Developmental Psychology in Beijing Normal University, China)*

Abstract

Polytomous item has been applied in more and more educational examinations, and researchers have developed several graded response models to describe this item type. But these models can only provide limited information and an inference about how well the student's integral ability is. Diagnostic assessment is a possible solution, where students may benefit from the diagnostic scoring skills. Based on Rule Space Model (RSM), one of the common diagnostic models, this paper discussed the method to extend the dichotomous rule space model to the polytomous model. RSM, based on the unidimensional IRT ability and person fit statistics, can be considered as a residual approach that employed a Q matrix. In the polytomous item situation, we can directly estimate the ability parameter with the parameter estimation techniques, but the person fit statistics need to be reconstructed,so we gave a extended zeta's formula and proved the properties that it should contain. We then employed the Monte Carlo method, compared the effect which the different attribute hierarchy and the scholastic response error has on the accuracy of the classification. The results showed that the accuracy was over 90% in various attribute hierarchy, and the scholastic response error had a significant impact on the classification accuracy. The accuracy decreased as the scholastic response error increased.

P02-23

Estimating raw items parameter in computerized adaptive testing

You, Xiaofeng (Foreign Language Teaching and Reseach Press, China)*

Abstract

Recently, with the development of computer technology and increasing needs of individual learning, CAT causes more and more concern. However, the security of adaptive test is also confronted with some new challenges. Lot's of factors, for example items used repeatedly and shared among the candidates ,affect the safety, effectiveness and fairness of the adaptive test .Therefore ,it is a very important issue in the research on computer adaptive test that build a large-scale, high-quality CAT item bank. At present, the items of CAT are generally constructed by experts, and then test takers are employed to acquire their responses to each item, estimation of item parameters and verification of FIT will be done. High-quality item and corresponding item parameters merge into the bank together. If raw items can be inserted in the process of CAT and the item parameters are estimated at the same time, it will be significant for the construction of CAT item bank. Now, it has not seen such kind of report in this field either in the national magazines or the foreign researches. Thus, the present research is to study the insertion of raw items and their parameters' estimation. A new kind of online calibrating is proposed and the formula of initial value during the stage of the iteration is determined. Simulation of Monte Carlo has been employed to estimate the parameters of raw items with the one-parameter Logistic and two-parameter Logistic models in the study and good result has been gained.

P03 08:30-10:00 Room105

P03-2

The effects of education level, occupational status and cultural capital on crystalized and fluid intelligence in adults

Coscarelli, Alessandra (University of Turin and University of Valle d'Aosta, Italy)
Balboni, Giulia (University of Valle d'Aosta, Italy)*
Cubelli, Roberto (University of Trento, Italy)

Abstract

In studies investigating the effects of Socio-Cultural Level (SCL) on intelligence tasks, SCL has been generally measured by the individual's socio-economic status (i.e., education level, occupational status, and income). Nevertheless, the cultural capital, an important indicator of SCL, has never been considered. The aim of this study is to explore the differential effects of educational level, occupational status and cultural capital on crystalized and fluid intelligence. The WAIS Vocabulary sub-test and the Raven's Advanced Progressive Matrices (APM) were administered to 80 Italian adults, aged from 30 to 45 years old, with a secondary school degree and a college degree. According to the questionnaire for cultural interests and the Italian scale of Occupational Status, participants were classified, respectively, with low and high cultural capital and with low and high occupational status. Results showed that APM is affected by educational level and occupational status but not by the cultural capital, whereas the Vocabulary sub-test is affected by all the three factors. Moreover, Vocabulary scores are influenced by the cultural capital in participants with low professional level and by educational level in participants with high professional level. In contrast, APM scores are influenced only by the educational level and only in participants with low professional level. In conclusion, performance on crystalized and fluid intelligence tests appeared to be differently affected by the different indicators of SCL.

P3-3

Validation of the Delinquency Reduction Outcome Profile (DROP) in a sample of incarcerated juveniles

Barbot, Baptiste (Child study center, Yale University, USA)*
Haeffel, Gerald (University of Notre Dame, USA)
Jon, Chapman (Court Support Services Division, State of Connecticut Judicial Branch, USA)
Elena, Grigorenko (Child study center, Department of Psychology, Department of Epidemiology & Public Health, Yale University, USA)

Abstract

The Delinquency Reduction Outcome Profile (DROP) is a new situational-judgment test (SJT) including 16 vignettes designed to measure social decision-making by delinquent youth. DROP combines both a typical SJT scoring method (distance to an "ideal" profile estimated reliably on the basis of ratings from experts), as well as a classic psychometric approach allowing the assessment of latent dimensions referring to distinct strategies used to approach the situations. We present the validation of this tool on a sample of 2033 participants recruited from juvenile detention centers. Both internal validity (multi-situation, multi-trait model) and external validity (using screening tools to assess mental health, posttraumatic stress disorder and social difficulties, as well as positive and negative behaviors observed in the detention centers) are described and discussed. Different forms of individual profiles on the latent dimensions are also presented and described in relation to their association with further outcomes (e.g., number of repeated admissions to detention, violation of detention rules, and risk of re-incarceration). In conclusion, we discuss practical implication of the DROP which can be used as an efficient prognostic tool to predict positive and negative outcomes, and to monitor changes in educational programs designed to improve social skills and reduce recidivism.

P03-4

Does frequent writing improve one's writing skills?

Han, Sangjun (Vantage Learning, Taiwan)
Edelblut, Paul (Vantage Learning /Summit IntelliMetric, Taiwan)
Chang, Kenneth (Summit IntelliMetric / Vantage Learning, Taiwan)*

Abstract

"Frequent writing improves one's writing skill" is a hypothesis that has long been accepted as truth without any solid evidence. But is it true? Has any teacher or student kept track of all their writing works with grades? What affects one's writing skills? What kinds of role do factors such as gender, language fluency, ethnicity, and other demographic data play in writing improvement? No one has ever been able to answer these questions due to various reasons. The most compelling reason was the difficulty of collecting writings from people of diverse background and demographic information and yet ensuring the objectivity of the scoring standards as if all of them had been graded by one expert scorer. This paper illustrates the effect of frequent writing practices on the overall quality of writing using the unique tool called MyAccess. MyAccess is an instructional writing tool that can grade any student's essays in a couple of seconds and return the holistic and five different domain scores as well as specific feedbacks on how a student can improve his/her essay. The massive amount of data from students in multiple grades and English proficiency levels with different genders and ethnic background has been collected and analyzed to test the long-time unchallenged hypothesis.

P03-5

A confirmatory cross – cultural investigation of the dimensionality of the Family Environment Scale

Charalampous, Kyriakos (Democritus University of Thrace, Greece)*
Kokkinos, Constantinos (Democritus University of Thrace, Greece)
Panayiotou, Georgia (University of Cyprus, Greece)

Abstract

The factorial structure of the Family Environment Scale (FES; Moos & Moos, 2002) has been examined by a number of studies. Results however, have been inconsistent in terms of the number of factors extracted. The purpose of the present study was to re-examine the first- and second-order factor structure of the FES by combining exploratory and confirmatory techniques, while testing simultaneously for the full model. The sample consisted of 409 individuals from Cyprus and Greece aged between 16

to 66 years. Results indicated that the initial 10-subscale structure of FES did not fit, for the most part, the data. A model with seven first- and three second-order factors was instead considered the best fit to the data. Overall, the findings indicate that the FES is both a valid and reliable instrument appropriate for use with samples of varying ages and cultures. While the results corroborate earlier findings which question the adequacy of the instrument's factor structure, a slightly different model, which supports the scale's general theoretical framework, is proposed. The present study is probably the first to apply an in-depth, step by step analysis and justification of the FES first- and second-order factor structure by simultaneously testing for a full second-order factor model.

P03-6

Psychometric evaluation of the Italian translation of the Alzheimer Functional Assessment Tool (AFAST).

De Vreese, Luc Pieter (Psychogeriatric Service, Health District, Modena, Italy)
Mantesso, Ulrico (ANFFAS Trentino Onlus, Italy)
De Bastiani, Elisa (ANFFAS Trentino Onlus, Italy)*
Gomiero, Tiziano (ANFFAS Trentino Onlus, Italy)

Abstract

AFAST is an informant-based questionnaire that quantifies on a six-to seven point ordinal scale, the degree of impairment or type of assistance required in six basic activities (toileting, dining,...,oral hygiene) and environmental awareness. Following a standardized process of forward and back translations to achieve a conceptual equivalence between the original English version and Italian translation (AFAST-I), its internal consistency and intra- and inter-rater reliabilities were tested on a sample of 61 adults with DS or other forms of ID with a mean (± SD) age of 53.4 (±7.71). All of the ID subjects attended day and residential services of Northern Italy. Cronbach's coefficient was .92 with all inter-item (mean .62; range: .43-.82) and total-item (mean: .82; range: .76-.90) correlations being above the .40 criterion, suggesting a high level of internally consistent reliability of AFAST-I. Inter-rater and intra-rater reliabilities too were excellent with ICC correlations computed on the AFAST-I sum scores of .96 and .94, respectively. The AFAST-I sum scores significantly correlated with both the IADL and BADL scores: Pearson correlations coefficients of -.69

P03-7

The development of the risk classification tool of adult probationer

Ding, Ching-Huei (National University of Tainan, Taiwan)*
Twu, Bor-Yaun (National University of Tainan, Taiwan)

Abstract

The development of economy is affected by the raising recidivism rate. Decreasing the recidivism rate can raise economy validly and protect people from risk. Recently, numerous researches focus on predicting recidivism, and the risk factors (dynamic, static factors). Assessment can provide highly accurate predictions of how offenders with similar characteristics might behave in the future. The purpose of this research is to develop a valid adult probationer risk classification tool. The factors assessed by the Risk Classification Tool of Adult Probationer will include the criminal history, financial, family and marital, leisure and recreation, companions, alcohol and drugs, life stresses, and antisocial personality problems. Some of them were assessed by the LSI-R (Andrews & Bonta, 1995) or SAQ (Loza, 1996). The Rasch model will be used to scale this tool. The risk assessment tool provides an improved classification of risk and identifies intervention targets for offenders during probation treatment. Key Words: recidivism rate, risk classification tool, probation, dynamic, static factors, Rasch model

P03-8

Effects of Imputation Methods for item non response on Cronbach's alpha

López-Jáuregui, Alicia (University of the Basque Country, Spain)
Elosua, Paula (University of the Basque Country, Spain)*

Abstract

Item non response is a practical problem in educational and psychological testing. The presence of "non available data" makes necessary to adopt decisions which can affect the estimation of the mathematical model. Thus far, the common practice has been to ignore observations with missing data (listwise deletion) or to apply naive methods like person mean imputation, despite the fact that these procedures can yield biased findings and loss of statistical power, specially with non-ignorable missingness. This study deals with the usefulness of one imputation method specifically adequate for such kind of data set, namely Response Function Imputation. Procedures to investigate the pattern of missingness are illustrated with data coming from the Eating Disorder Inventory-3 questionnaire; two imputation methods were applied and their performance in the estimation of reliability index were compared.

P03-9

Development of indigenous scale to measure job autonomy

Fida, Kashif M. (Forman Christian College (A Chartered University), Lahore, Pakistan)*
Najam, Najma (Karakrum International University, Gilgit, Pakistan)

Abstract

Development of indigenous scale to measure job autonomy (JA) among employees of private and privatized organizations. Job autonomy is the extent to which employees have major say in scheduling their work, selecting the equipment they use and deciding the procedures to be followed (Wagner & Hollenbeck, 1998). Focus group technique was used to explore Job Autonomy among 36 employees of private and privatized organizations. Discussions from focus groups were utilized in construction of 28 statements. Likert scale (5-point) was adopted to collect the responses. 236 employees were participated in this research, from private and privatized organizations. Reliability analysis showed significant Cronbach's alpha=0.72. Correlation coefficient was computed among the sub-areas of JA and found with relatively low correlations (p>0.01) showing significant discriminant validity. Factor analysis yields 7 sub-areas of J.A. i-e. 1. Job functioning 2. Scheduling, 3. Decision making, 4. Physical environment, 5. Gender, 6. Social interactions, and 7. Dress code. JA Scale is indigenous, reasonably reliable and valid for the assessment of JA of employees of private and privatized organizations. This scale may help the professionals, decision makers, authorities and employer to find out the autonomy of job, its requirements, and intensity among employees of Pakistan.

P03-10

Narrow personality traits, work motivation and sales performance

Iversen, Ole (Norwegian School of Management BI, Norway)*
Hatlem, Morten (Schibsted ASA, Norway)

Abstract

In today's competitive marketplace an effective sales force is an asset for a company and there has been an increasing awareness of the importance of personality as a predictor of job success during the last couple of decades. This study investigates the relationship between narrow personality traits, motivation and sales performance for 75 salespeople working in the media sector in Scandinavia. Personality data was collected by the OPQ 32, whereas the MQ was used to collect motivational data. Sales performance was measured objectively by actual sales results. Interestingly neither Social Confident, Contentious nor Achieving were found to correlate with sales performance. Moreover, four other personality traits (Controlling, Conventional, Tough Minded and Rule Following) were found to explain 15% of the variation in sales performance. Including two motivational factors in the model (fear of failure and flexible) increased the explained variance in sales performance to 24.2%, No relationship was found between experience (length in role) and sales performance. The MQ related extrinsic motivation domains did not explain sales performance, indicating that many companies may waste time and money on individual bonus systems. The findings indicate that incremental validity can be achieved by combining narrow personality measures with motivational measures in a selection process; furthermore the findings suggest that personality and motivation are more important than experience for sales personnel.

P03-11

Chinese Linguistic Inquiry and Word Count (CLIWC): Developing and applying a linguistic analysis program in cultural context

Hui, Natalie H. H. (Hong Kong SAR, China)
Lam, Ben C. P. (Hong Kong Polytechnic University, Hong Kong SAR, China)*

Abstract

Our use of words can reveal important information about our personality, psychological states and social connections to the world. The Linguistic Inquiry and Word Count (LIWC) program (see www.liwc.net for more information) was originally developed by Pennebaker and colleagues to conduct text analysis of narratives on a word-by-word basis. Since its release, LIWC has been widely used by research psychologists in the sub-fields of personality, social and health psychology to investigate a wide range of psychological phenomenon. Despite of numerous empirical evidences gathered through English-speaking samples, limited research has been conducted using Chinese-speaking sample due to the lack of a comparable version of LIWC in Chinese. In expanding our research toolbox, we developed the Chinese version of LIWC (CLIWC) that enables us to perform linguistic analysis with Chinese-speaking populations. We will present the newly developed CLIWC and its development. Moreover, we will present some preliminary evidences validating CLIWC as a useful research tool using writing samples of individuals with different levels of psychological distress, viz. depression and suicidal ideation. Our findings showed that respondents' level of depression was positively associated with percentages of first person singulars (e.g., I) and negative emotional words (e.g., sad). Furthermore, individuals' level of suicidal ideation was positively related to percentage of death words (e.g., die). Other emic word categories will also be examined. Lastly, the use of CLIWC in future research as a low-cost and unobtrusive analytical tool in studying the psyche of Chinese in written materials will be discussed.

P03-12

Prediction of conscientious behavior using implicit and explicit measures of trait conscientiousness

Laurentiu P., Maricuoiu (West University of Timioara, România)*
Irina, Macsinga (West University of Timioara, România)
Delia, Virga (West University of Timioara, România)
Silvia, Rusu (West University of Timioara, România)
Cheng, Clara M. (American University, USA)
Sava, Florin A. (West University of Timioara, România)

Abstract

According to the Reflective-Impulsive Model (Strack & Deutsch,2004) observed behavior is a function of reflective (or explicit) and impulsive (or implicit) processes. The present research investigates the predictive power of explicit and implicit measures of trait conscientiousness on conscientious behavior. Participants were 94 students. Behavioral measures of conscientiousness were: participants' grade-point average (GPA), an observer evaluation of participants' behavior in a group task; and lateness of attendance at the study. These measures were combined in a single factorial score. For assessing explicit conscientiousness we used three questionnaires that measured the Big Five model: IPIP (Goldberg, 1993), NEO-FFI (Costa & McRae, 1992), DECAS (Sava, 2008). The implicit measures of conscientiousness were assessed with the Implicit Association Test (Greenwald, McGhee & Schwartz, 1998) and the Identification Misattribution Procedure (Sava et. al, 2009). We conducted an exploratory factor analysis for examining the latent structure of the implicit and explicit scores. Results indicated a two-factor solution. We obtained significant correlations between the behavioral, explicit and implicit factorial scores of conscientiousness. Then, we conducted a multiple regression analysis for evaluating the predictive efficiency of implicit and explicit measures on the behavioral factor of conscientiousness. Results indicated that explicit and implicit measures contribute in a cumulative manner in explaining the variance of the conscientious behavior. This result supports the idea that implicit measures can provide better understanding and better prediction of actual behavior.

P03-13

Expatriate job performance: Development of a criterion measure for use in the expatriate support and management practice domain

Lee, Leanda (Monash University and University of Saint Joseph Macau, China)*
Care, Esther (Assessment Research Centre, University of Melbourne, Australia)

Abstract

Based upon Campbell's (1990) multi-factorial model of job performance the Expatriate Performance Scale was developed to measure the components of expatriate performance. Item creation for the scale was informed by job performance theory applicable to both expatriate and managerial job performance with further exploration of the construct undertaken through content analysis of data from semi-structured interviews with 20 expatriate employees. The scale development comprised a review of items and sorting by subject matter experts prior to a pilot study, the data from which were subjected to item analysis. The modified scale comprising 48 items was then administered to a sample of 106 Australian expatriate employees in the Special Administrative Regions of China. These data were subjected to reliability analysis, and examined for factorial structure using Principal Components Analysis with oblimin rotation. The scale development and validation process resulted in 32 items measuring an amended model of expatriate performance with six components rather than Campbell's original eight. This study contributes to expatriate performance theory by distinguishing between components and predictors of job performance. The development of a scale based upon solid deductive and inductive foundations also nullifies the criticism often applied to the domain that it is purely conceptual in nature. In practical terms this scale has potential to offer a multi-dimensional measure of expatriate performance to determine to what extent various organizational support mechanisms and individual level variables predict variance in each of the six components of performance.

P03-14

Bayesian methods for studying DIF in a quality of life model

Lozano, Luis M. (University of Granada, Spain)*
Monsalve, Carolina (University of Granada, Spain)
Perez-llantada, Carmen (UNED)
Lopez de la llave, Andres (UNED)

Abstract

The perceived quality of life is a very important variable in the way the people interacts with the environment. The WHO defines the Quality of Life as individual perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. Enjoy a good quality of life in elderly is really important because it means have a good functional skills and a general satisfaction, with physical and psychological welfare. In this sense, the Spanish Dependency Law tries to encourage that the elder people keep their own autonomy, bringing some supplies to reach it. The need of creating good questionnaires to evaluate the perceived quality of life in an accurate and valid way, so the supplies can be provided by the administration in a fair way, is high. The Bayesian framework will provide different techniques that can allow to the test developers to estimate the reliability, validity, DIF of a questionnaire The aim of this work is to present the Bayesian way, using ordinal logistic regression, to flag items that can be biased, using a perceived quality of life questionnaire as a field to develop this technique. This work has been founded with the project SEJ2006-13009 of the Ministry of education and Science and the project P07HUM-02529 of excellence of the Junta de Andalucía

P03-15

Information and everyday life: A French assessment of adult literacy

Rocher, Thierry (Ministry of Education (DEPP) / Université Paris Ouest Nanterre La Défense, PPCC, EA 4431, France)
Megherbi, Hakima (Université Paris 13, UFR LSHS, UTRPP, EA 3413, France)*

Abstract

Following methodological problems that appeared in France in the IALS survey (International Adult Literacy Survey), the French Institute of Statistics (INSEE), in collaboration with the Ministry of Education (DEPP) and other institutions, has conducted in 2004 a French national survey called 'Information et Vie Quotidienne' ("Information in Everyday Life"). The survey aims to measure literacy performance in everyday life and has been conducted in 2004 with about 10,000 adults aged from 18-to-65 year-olds. Some methodological aspects of the survey are presented. A particular attention was given to the relation between the interviewer and the assessed person, to guarantee motivation and implication. Items from IALS have been used and it revealed that instead of 40 %, the proportion of French adults at the lowest level of Prose Literacy seems to be about 15 %. A two-stage design with a routing test was implemented in order to distinguish between illiterate people and good readers. Given this adaptive design, the question of the calculation of a global score calculation was also investigated.

P03-16

Validating the insights discovery preference evaluator: A global project

Okonkwo, Judith (The Business Psychology Centre/University of Westminster, United Kingdom)*
Benton, Stephen (The Business Psychology Centre)
van Erkom Schurink, Corine
Desson, Stewart
Brenstein, Elke

Abstract

To provide evidence of the psychometric probity of a tool that measures Jungian preferences – the Insights Discovery Preference Evaluator (IDPE) - according to current EFPA guidelines; while meeting the research needs of a global learning and development organization and its international clients. Methods: A variety of samples – convenience, client sourced, student populations etc (n= 34 to 519,467) were utilized to carry out standard tests for validity and reliability, which included Test-retest, Cronbach alpha, Item analysis, Factor Analysis, Construct validity and various Criterion Validity studies. Results: Objectives were achieved with reliability and validity further established for the IDPE, e.g., Cronbach alpha coefficients ranged from 0.917 – 0.932 etc. Conclusion: This case study explored the validation process of a web based psychometric tool. As we contend with competing stakeholder requirements and the growing influence of the internet on psychometrics, how do we adjust our research processes to accommodate the peculiar situations that arise? This presentation details a validation process that embraces both academic and commercial needs, lending itself to further business opportunities while maintaining the scientific rigor required.

P03-17

Development of a cross-national community flourishing questionnaire

So, Timothy (University of Cambridge, Wellbeing Institute and Global Chinese Positive Psychology Association, UK)*
Huppert, Felicia (University of Cambridge, Wellbeing Institute, UK)
Willoughby, Andrew (Global Community Well-being)

Abstract

Survey designers are becoming increasingly interested in measuring subjective aspects of well-being. To date however most subjective measures of well-being are concerned with the well-being of individuals. There is a need to also measure the well-being of communities, which can be a local area, a school, a business or other organization. We have developed a Community Flourishing Questionnaire (CFQ) with the aim of describing the extent to which a community is perceived to be flourishing at a particular time, and how this relates to a variety of socio-economic and other characteristics. The CFQ investigates a number of dimensions, including community structures, community processes, community values, community cohesion and participation in community activities. The CFQ has been developed through extensive literature review and incorporates a number of existing scales and items, together with newly developed items to measure constructs which were not sufficiently represented in existing measures. After cognitive testing in the UK and HK, and consultation with international experts, the CFQ was piloted on a representative sample of the population in the UK and HK. Analysis of the psychometric properties of the questionnaire suggest that with appropriate refinement it is a reliable and valid measure for use in cross-national studies. The CFQ can also be used as a measure of change over time, or following community-level interventions, aimed at increasing the perception of how well a community is functioning.

P03-18

Cross-cultural studies of leadership using questionaires and a 360° approach: To which extent are the differences important and meaningful?

Vrignaud, Pierre (Université Paris Ouest Nanterre La Défense, France)*
Ket de Vries, Manfred (INSEAD)
Florent, Elizabeth (INSEAD)

Abstract

The leadership questionnaires assess the main dimensions that explain the difference of efficacy in managing a team. They are often used in the framework of a 360° approach : the subject him/herself responds to the questions, and the same questions are also answered by several (3 to 10) observers who know him/her well. At the INSEAD Global Leadership Centre (IGLC), we constructed and validated a leadership questionnaire: the Global Executive Leadership Inventory (Ket de Vries, Vrignaud, & Florent-Treacy, 2004). Since this publication, we gathered the data of respondents from many different countries. Our database includes more than 6,500 leaders and 50,000 respondents from more than 100 nationalities. The data collected by a 360° approach have a hierarchical structure: the observers are embedded in a self. The importance of this structure has been generally underlined but seldom completely taken into consideration. Multilevel modeling has been developed to analyze reliably such data by estimating the variability pertaining to the different level (Bryk & Raudenbush, 2002 ; Goldstein, 2003). For our cross-national analyzes, we integrated the countries as an upper hierarchical level. The analysis of these results allows to wonder about the real importance of the between countries variability and of the possibility to bring out robust and interpretable clusters merging the different countries into higher level units. We try to explain the reasons of these difficulties by discussing several psychometrical issues bound to the questionnaire reliability in the context of a 360° and by analyzing the leaders' background in relation with their nationalities.

P03-19

Direct estimation of the correlation between latent traits within the multidimensional IRT framework

Wan, Shih-Ting (The Graduate Institute of Educational Measurement and Statistics , the National Taichung University, Taiwan)*
Wang, Wen-Chung (Department of Educational Psychology, Counselling and Learning Needs,the Hong Kong Instittute of Education, Hong Kong)
Shih, Ching-Lin (The Graduate Institute of Educational Measurement and Statistics ,the National Taichung University, Taiwan)

Abstract

In educational and psychological studies, researchers are often requested to reporting effect sizes, such as the strength of relationship between latent traits. The estimation of the correlation between latent traits will be attenuated if measurement error exits but not taken into account. In the framework of classical testing theory (CTT), dis-attenuation formula has been developed to adjust the estimation. Nowadays, item response theory (IRT) has been widely used to replace CTT in analyzing educational and psychological test data. How to estimate accurately the correlation between latent traits within the IRT framework becomes important. Previously studies (e.g., Wang, 200; Wang, Chen, & Cheng, 2004) have shown how to apply multidimensional Rasch models to estimate the correlation between latent traits directly. In this study, we further develop a multidimensional IRT model that contains multidimensional Rasch models as special cases and demonstrate how to apply it to empirical data. To be general, the item response functions in the new multidimensional model can follow the 1-parameter (Rasch), 2-parameter, or the 3-parameter logistic model for dichotomous items, and the (generalized) partial credit model, the rating scale model, or the graded response model for polytomous items. In addition, testlet response modeling is also possible. Parameters in the new model can be estimated with the computer software WinBUGS. A series of simulations were conducted to assess parameter recovery. The 2007 TIMSS data were analyzed. The simulation results showed that the correlation between latent traits as well as parameters can be recovered accurately with WinBUGS. With respect to the empirical data set, the multidimensional model yielded a much higher correlation than traditional unidimensional approaches where measurement errors in person measures are ignored. In conclusion, the new multidimensional model is very flexible and can be used to estimate the correlation between latent traits.

P03-20

IRT-based unfolding models and their extensions

Wu, Shiu-Lien (National Chung Cheng University, Taiwan)*
Wang, Wen-Chung (The Hong Kong Institute of Education, Hong Kong SAR, China)

Abstract

IRT-Based Unfolding Models and Their Extensions Shiu-Lien Wu National Chung Cheng University, Taiwan Wen-Chung Wang The Hong Kong Institute of Education Background and Objectives Several IRT-based unfolding models have been developed to fit responses of attitude items, including the hyperbolic cosine model (Andrich, 1988, 1996), the PARELLA model (Hoijtink, 1990) and the generalized graded unfolding model (Roberts, Donoghue, & Laughlin, 2000, 2002). Among these models, the generalized graded unfolding model (GGUM) appears to draw more attention of researchers and practitioners, because it is general and its freeware for parameter estimation. This study aims at extending the GGUM to further strengthen its applications in general settings. Methods The GGUM can be extended in twofold. First, the threshold parameters can be treated as random-effects rather than fixed-effects to depict the randomness in subjective judgment when participants respond to Likert items. The resulting model is called the random-threshold generalized graded unfolding model (RTGGUM). Second, the unidimensional GGUM can be generalized to be multidimensional, which is called the multidimensional generalized graded unfolding model (MGGUM). The new model takes into account the interrelationship between latent traits to increase measurement precision and yields a direct estimation of the interrelationship which is free from the attenuation of measurement error in person measures. A series of simulation studies were conducted to assess parameter recovery using the computer software WinBUGS. Empirical examples were also given to demonstrate implication and application of the new models. Results and Conclusion The simulation results indicated a good parameter recovery for the new models. In general, the fixed-effect parameters were easier to recover than the random-effect parameters. Empirical data analyses suggested that the RTGGUM had a better fit than the GGUM, indicating randomness in subjective judgment among participants, and that the MGGUM yielded a higher test reliability and a higher correlation between latent trait than the GGUM. In conclusion, the two new models have been successfully developed and they are more flexible and efficient than the traditional GGUM. Keywords: item response theory, generalized graded unfolding model, Likert item, random effect, multidimensional model.

P03-21

What is effective English teaching: Perceptions from New Zealand high school students

Brown, Gavin Thomas Lumsden (The Hong Kong Institute of Education, Hong Kong SAR, China)*
Yu, Keling (The Hong Kong Institute of Education, Hong Kong SAR, China)

Abstract

This study investigated 684 students of year 11 and year 12 in the secondary schools in New Zealand and analyzed their reading performance, reading attitudes and their evaluating English teaching data. The reading performance was measured using the Assessment Tools for Teaching and Learning version 4 (asTTle V4) test system (Hattie, Brown, Keegan, MacKay, Irving, Cutforth, et al., 2004). Relating to reading attitudes, students answered six items drawn from the National Education Monitoring Project and the results showed that how much they like and how good they were at the English reading (Otunuku & Brown, 2007). About students' evaluating teaching, the 10 items in questionnaire to investigate were adapted from Students Evaluating Accomplished Teaching – Mathematics (SEAT-M) (Irving, 2004). This study focused more on the relationship between students' evaluation of teaching and students' reading performance. We used SPSS to analyze the factor of reading attitude and students' evaluation of teaching. Then, with Lisrel, we conduct CFA and SEM to confirm the relationship. Findings showed that English teacher's subject knowledge and their instruction strategies would be predictor of student reading attitude and reading performance. However, teacher contacting parents would negatively affect student reading attitude and performance.

P03-22

Testing the dimensionality of Urban Identity Scale in two cities of China

Zhuang, Chunping (Institute of Psychology, Chinese Academy of Sciences, China)*
Huang, Fei (Institute of Psychology, Chinese Academy of Sciences, China)
Zhang, Jianxin (Institute of Psychology, Chinese Academy of Sciences, China)

Abstract

Urban Identity Scale established by Lalli (1992)is a widely used instrument to measure the place identity. To test the applicability of the scale in China, we made this investigation to examine the stability of the factor structure and other psychometric properties of Urban Identity Scale. An online survey was taken with a sampling of 392 participants living in Beijing (N=133) and Shanghai (N=259). These data supported a five-factor model consisting of external evaluation, general attachment, continuity with personal past, perception of familiarity, commitment. All of those dimensions were in line with the original structure of the scale. Owing to the translation reasons and cultural differences, we set the errors of four items correlated. A series of multiple-groups confirmatory factor analysis was conducted to assess the measurement invariance of the Urban Identity Scale across the two cities. The latent factor means were compared and no differences were found between the two cities. Results indicated that the Urban Identity Scale was applicable in China.

P04 10:30-12:00 Room105

P04-1

Utilising Item Response Theory (IRT) and Differential Item Functioning (DIF) for adaptation of a perceived control scale between cultures and languages

Ertubey, Candan (University of Bedfordshire, UK)*

Abstract

Cross-cultural psychology has been trying to address the issue of languages as a term of bias during the investigation of different cultures. This enables us to look at the effect of culture independent of the effect of language. The present study looks into the use of IRT as a modern psychometric approach in the event of adaptation of a scale between two languages (English & Turkish). The scale used in the study was the CAMI scale (Skinner, Baltes, Chapman, 1998), which measures perceived control using 12 subscales with Likert style items. The aim of the study was to use IRT and DIF to adapt the CAMI scale to English and Turkish languages and thus make a comparison between two cultures. This will enable us to investigate similarities and differences between English and Turkish population in relation to perceived control. To investigate the differences between item parameters, item discrimination "a" and difficulties "b" were calculated for subscales. The analysis was conducted via the MULTILOG program. A total of 813 pupils (English N= 365 and Turkish N=448) between 14 and 18 years old answered the scale in their own language. The results found differences between the two groups of students on 4/12 subscales when the DIF was tested with X2 .The subscales were then analyzed at the item level and the differences substantiated. The results can be interpreted as showing that IRT and, specifically, DIF are useful ways to investigate the differences between cultures subscale and item level.

P04-2

Psychometric properties of SIMTEST: A computerized test of foreign language proficiency

García-Rueda, Rebeca (Universitat Autònoma de Barcelona, Spain)*
Doval, Eduardo (Universitat Autònoma de Barcelona, Spain)*
Viladrich, Maria Carme (Universitat Autònoma de Barcelona, Spain)
Sumbling, Mick (Universitat Autònoma de Barcelona, Spain)
Riera Grau, Laura (Universitat Autònoma de Barcelona, Spain)
Sanz, Pablo (Universitat Autònoma de Barcelona, Spain)

Abstract

SIMTEST is an on-line, multi-component test of foreign language proficiency; designed and developed at the Universitat Aut;noma de Barcelona (UAB). It has been used to place students on English as a Foreign Language (EFL) courses, and to certify their level of proficiency in terms of the Common Reference Levels of the 'Common European Framework of Reference for Languages: learning, teaching, assessment'. SIMTEST consists of an initial C-test component: an integrative written test based on the concept of reduced redundancy and comprising 4 texts randomly selected from the item bank. While each text has an intact title and opening and closing sentences, the central part of each consists of 25 'damaged' words to be 'repaired' by the candidate. C-tests were explicitly designed to measure global language proficiency and have been selected for inclusion on SIMTEST on the basis of the similarity of their psychometric properties. The second component of SIMTEST is a computerized adaptive test (CAT) testing knowledge of grammar, vocabulary and linguistic functions with multiple-choice items. A third SIMTEST component, used in certification, consists of a second CAT, also in multiple-choice format, testing foreign language listening comprehension. In this study we report evidence of the psychometric properties of SIMTEST based on data from some 3000 students at the UAB and the Universidad Complutense de Madrid gathered since 2001. Results deal with item calibration, internal structure, internal consistency, sensitivity to change, in relation to English level based on both expert judgment and outcomes on the Oxford Quick Placement Test (QPT).

P04-3

Test adaptation : How to deal with practical issues

Gillet, Isabelle (Editions Hogrefe France, France)*

Abstract

There is a rising demand for the same metric in different languages and different cultures. Along with this demand goes of course the need to be assured that the test 'is the same' with no bias. But the question of cultural equivalence in test adaptation has no easy answer. Despite the different guidelines on how to achieve it, the fact is that we have not yet reached an ideal solution. With concrete examples from test adaptation processes, of personality questionnaires and cognitive batteries, a model for adaptation will be presented and discussed. As we cannot assess personality traits directly; we can only infer them from behaviour; the underlying construct may generalize across cultures but its manifestation in behaviour will vary because culture impacts on behavioural expression. Moreover personality test items are expressed in language which itself is full of cultural, social and mental representations. For cognitive tests, even figural items (symbols and images) does not free the test-taker from culturally based associations. The work of adapting a test from the source language to the target language progresses through the items. The items represent the way the construct is expressed in the source culture and we cannot expect this to remain constant across cultures

P04-4

Comparing UIRT, MIRT, and HIRT based on model fitting and parameter recovery

Kuo, Bor-Chen (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)*
Hsieh, Tien-Yu (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)
Cheng, Chien-Ming (National Academy for Educational Research Preparatory Office, Taiwan)

Abstract

The hierarchical item response (HIRT) model was proposed by de la Torre and Song (2009) to model one overall score and domain scores concurrently. de la Torre and Song (2009) showed that the uni-dimensional item response theory (UIRT) model is the special case of the HIRT model. In this study, the performances of UIRT, MIRT, and HIRT are compared based on the model fitting indices and parameter estimation accuracy with simulated and real data. Four indices are considered: Akaike's information coefficient (AIC), Bayesian information coefficient (BIC), deviance information coefficient (DIC), and posterior predictive model checks (PPMC). And the root mean square error (RMSE) is used for measuring the parameter recovery. The following factors are considered in the experiments: 1. Test structures: between-item test and within-item test 2. Sample sizes: 1000 and 4000 examinees, 3. Test lengths: 20, 40, and 80 items. Experimental result shows that the HIRT model can fit the simulation data generated by the UIRT model and the MIRT model. Keyword: HIRT; UIRT; MIRT; model selection; AIC; BIC; DIC; PPMC.

P04-5

Application of many-facet model on the process-assessment and product-assessment of creativity test

Hsu, Chun-Yu (National Taiwan Normal University, Department of Educational Psychology and Counseling, Taiwan)*
Chen, Po-Hsi (National Taiwan Normal University, Department of Educational Psychology and Counseling, Taiwan)

Abstract

The purpose of the study is to analyze the rater effect and criterion effect of the process-assessment and product-assessment of the creativity test. Two hundreds and seventeen university students were asked to answer a creativity test. In this test, subjects were asked to choose 3 geons of 36 geons to design a creative furniture. Two of three well-trained raters were asked to score the process and product of the creativity test using seven scoring criterions Many-facet Rasch model was used to analyze the data. Results showed that the multidimensional many-facet model fitted the data better than the unidimensional many-facet model, both in process assessment and product assessment. This indicated that different latent traits are involved in the process and product assessment of the creativity test. Results of the rater effect and criterion effect indicated both process and product assessment were not influenced by raters, however, they were influenced by some criterions. The results of this research will be applied on the automatic scoring module in the creativity test.

P04-7

The extension and application of the testlet response model for ability-based guessing

Lin, Yi-Hung (University of California, Berkeley, USA)*
Wilson, Mark (University of California, Berkeley, USA)

Abstract

Guessing is commonly found in multiple-choice (MC) items. The success of guessing has been found to be related to ability. Recently, the one-parameter logistic model for ability-based guessing has been proposed in 2006. A testlet is a set of items that share a common stimulus. Due to the sharing of common stimulus, items within a testlet may be locally dependent. The Rasch testlet model was proposed in 2005 to model the local dependence within a testlet. Both the ability-based guessing and testlet effects can be found together in MC items. To integrate the two models into one, the testlet response model for ability-based guessing was also suggested in 2009. However, the three models assumed that the discrimination parameter is a constant. This may be too restrictive in practice. To relax this restriction, a discrimination parameter is integrated into TRM-AG and formed a more flexible model, the generalized testlet response model for ability-based guessing (GTRM-AG). This study includes a simulation and an empirical study: The former was performed to examine the efficiency of parameter recovery for GTRM-AG, and the latter was conducted to demonstrate the application of GTRM-AG and compare the model fit of GTRM-AG and 3PLM. Roughly speaking, the parameters of the GTRM-AG had been recovered well, and GTRM-AG is the better fitting model if the testlet effect and ability-based guessing have substantial influences. When these effects exist but ignored, examinees would have a rather inappropriate ranking. This is a serious issue for test fairness, especially the large-scale entrance examination.

P04-8

Bayesian analysis to identify norming groups in the MBI questionnaire

De La Fuente, Emilia I. (University of Granada, Spain)
Lozano, Luis M. (University of Granada, Spain)*
Canadas, Guiliermo A. (University of Granada, Spain)
Martin, Maria (University of Granada, Spain)

Abstract

Norming is one of the steps that a psychologist hast to develop when is creating a questionnaire, but usually nobody brings the adequate time to this process. Wrong norms mean that the consequences that may derive from the questionnaire are erroneous, so the objective that the researcher wants when applies the test is not reached. Sometimes different norms may be make for different groups (males vs. females, for different levels of education, profession...). To know if these differences have to be make the researchers usually use the Analysis of Variance. From the frequentist point of view to use this technique the data have to meet some parametric assumptions. The Bayesian perspective brings similar results than the frequentist when the sample is big, but brings more accurate results when it is small. Other point is that the Bayesian perspective allows the researcher to use initial information. In the Frequentist framework the conclusions are if the differences are significant or not, but the Bayesian brings the probability of a mean of being bigger than the other, so the researchers can decide if must produce different norms for different groups. The aim of this work is to compare the two perspectives with different sample sizes in the maslach Burnout Inventory (M.B.I.).

P04-9

Bayesian ordinal logistic regression to flag biased items in a burnout questionnaire

De La Fuente, Emilia I. (University of Granada, Spain)
Lozano, Luis M. (University of Granada, Spain)*
Canadas, Guiliermo A. (University of Granada, Spain)
Concepcion San Luis (UNED)

Abstract

The burnout syndrome is characterized by the lack of motivation, disinterest, internal unrest or job dissatisfaction that seems to affect more or less to an important professional group. From a psychosocial perspective, the "Burnout" is conceptualized as a process that involves attitudinal-cognitive elements -reduced personal accomplishment at work-, emotional,-emotional exhaustion-, and attitudinal – depersonalization-. The Maslach Burnout Inventory (MBI) has become the most used questionnaire to assess burnout in more recent years. The logistic regression is one of the most flexible techniques to flag biased items of a questionnaire. In psychology lot of works recommend the use of polytomous items because they improve the psychometric properties (reliability and validity) of the test. In these kind of items the logistic regression is not correct, so different authors recommend the use of the ordinal logistic regression. It has the same modeling strategy than the logistic regression, it is very easy to implement… The aim of this work is to make a psychometric valuation of the M.B.I. questionnaire, from the DIF framework. To reach this objective the ordinal logistic regression is used. This work has been founded with the project SEJ2006-13009 of the Ministry of education and Science and the project P07HUM-02529 of excellence of the Junta de Andalucía

P04-10

Using IRT modeling with situational judgment tests: Items vs. testlets

Li, Tao (Hogrefe Ltd, UK)*

Abstract

Situational judgment tests (SJTs) have gained increased popularity in personnel selection research and practice. SJTs present test takers with work-related scenarios and possible responses to the situations. In SJTs, test items are often grouped into bundles, or item sets, centered around a common stimulus (i.e. scenario). As such, SJT items are context-dependent. The item response that emerges out of the test format is the interaction between a person and a simulated situational context. The context-dependent nature of SJTs means that conventional one-factor measurement models are frequently violated by local item dependency. Fitting item response theory (IRT) model with item level data often leads to biased item parameter and information (reliability) estimation. Employing the concept of testlet response theory, the study aims to model the dependencies of SJTs items. Data from the Leadership Judgment Indicator (LJI, Hogrefe) are used to demonstrate the approach. Practical implications for scoring and analyzing SJTs are discussed.

P04-11

Application of association rules in cognitive diagnosis

Luo, Ling-Yun (Jiangxi Normal University, China)
Ding, Shu-Liang (Jiangxi Normal University, China)*
Gan, Deng-Wen (Jiangxi Normal University, China)

Abstract

The test not only serves evaluative purposes, but also offers valuable information regarding each examinee's educational needs. The purpose of Cognitive Diagnosis is to identify which attributes are mastered by a test taker and which ones are not. Various models have been proposed for cognitive diagnosis, including the Rule Space Model(RSM) (Tatsuoka, 1983), Attribute Hierarchy Model (AHM)(Leighton,2004),DINA model (Junker & Sijtsma, 2001) and the NIDA model (Maris, 1999). According to the identified hierarchical ordering of attributes, AHM assigns the observed examinee response patterns to the expected examinee response patterns, and estimates examinees' cognitive states. Q-Matrix identified by experts might not exactly reflect students' Knowledge Structure, so Q-Matrix misspecification may reduce the classification correct rate.(Rupp and Jonathan Tempin,2007). A problem might be encountered --- inconsistency of hierarchical ordering of attributes in AHM and a method based on association rules(ARs) in data mining is proposed to solve it and an algorithm is developed to automatically identify attributes for each item in AHM in this paper. The phenomenon of negative association rules is eliminated by using Correlation Coefficient to automatically identify attributes for each item in AHM and an index to determine the quality of hierarchical ordering of attributes is defined. Comparison with other strategies, such as Support Vector Machine(SVM),Neural Network,the result obtained by using ARs indicates that ARs is more stable , more accurate and less time of calculation.

P04-12

Using Multidimensional Scaling and Fuzzy Algorithm to assess mnemonic abilities during late stages of ontogenesis

Molchanov, Alexander (Moscow State University of Medicine and Dentistry, Russia)*
Molchanov, Kirill

Abstract

Multidimensional scaling (MDS) and fuzzy decision making can be used for assessing qualitative and quantitative changes of mnemonic abilities during late stages of ontogenesis. Individual differences Euclidean distance model of MDS (INDSCAL) make it possible to determine the number, "weight" of most significant properties of stimulus material during images memorizing. Using MDS, judgments of the dissimilarity of all stimulus pairs, collected from all subjects, are represented as two types of psychological spaces: as stimulus configuration space and space of subjects. Comparison of mnemonic abilities estimations performed by mean of psychological tests and properties of both types of psychological spaces has allowed to interpret properties of psychological spaces in terms of (1) inborn mnemonic abilities mechanisms, and also generated during lifetime (2) operational mechanisms (processing methods), and (3) control mechanisms (mostly attributing to personality characteristics) of mnemonic abilities. It has been established, that various ways of storing and particular features of mnemonic abilities are reflected in properties of psychological spaces. Age differences of the psychological spaces have been determined for groups of 50-59; 60-69; 70-79 years old subjects. For development of the mnemonic abilities individual correction plans the data are processed by fuzzy algorithm.

P04-13

Geotrends in testing – Coping with regional technology gaps: A case study of the validation of a leadership inventory

Okonkwo, Judith (The Business Psychology Centre/University of Westminster, UK)*
Benton, Stephen (The Business Psychology Centre, UK)
Brenstein, Elke
Desson, Stewart

Abstract

To review a global project aimed at providing evidence of the psychometric probity of a leadership tool that measures competencies in various formats (self-report and 360 feedback) - the Insights Navigator Transformational Leadership – guided by EFPA requirements; while meeting the research needs of a global learning and development organization and its international clients. We will discuss a number of methods, and results from the successful criterion validity studies on leadership in project management will be shared. This presentation will detail some of the limitations encountered in attempting to validate an instrument on a global scale using web based technology, we will highlight our experience in Uganda where we had to suspend our plans to collaborate with Kampala International University and future plans in that regard. We will also demonstrate the benefits of pursuing validation on a semi-commercial route highlighting the work done with the Project Management Institute around the world, proffering a knowledge sharing model that benefited all stakeholders - professional body, academic institution and the company. In the end we recommend that a validation process embraces both academic and commercial needs; lending itself to further business opportunities while maintaining the scientific rigor required.

P04-14

Applying IRT to the Advanced Progressive Matrices –Short Form

Primi, Caterina (Department of Psychology, University of Florence, Italy)*
Galli, Silvia (Department of Psychology, University of Florence, Italy)
Ciancaleoni, Matteo (Department of Psychology, University of Florence, Italy)
Chiesi, Francesca (Department of Psychology, University of Florence, Italy)

Abstract

Arthur and Day (1994) developed a 12-item short form of the Raven Advanced Progressive Matrices (APM-SF) that provides a sound assessment of general intelligence in shorter time frame. Since one of the psychometric properties of the long form is the increasing order of items difficulty level, the aim of the present study is to test this property applying the Rasch Simple Logistic Model (RSLM) to the APM-SF. The test was administered to a sample of 758 students (mean age= 22.5, SD=4.2; M= 53%). Confirmative factor analysis attested the monodimensionality of the test. Specifically the chi square/df ratio was 1.73; moreover the CFI and the TLI were both .98 and the RMSEA was .03. The items fit statistics were calculated through Winsteps (Linacre, 2005). The fit statistics revealed a good fit of each item to the model. Each item showed mean square infit statistics within a 0.7 and 1.3 range (Bond & Fox, 2007). The item difficulty measures covered a range between -2.23 ± 13 and 1.43± 09 logits. Some items difficult levels were not consistent with the way in which items have been selected to increase in difficulty. The application of RSLM does not confirm the progressive order of the item difficult level, and it suggests the possibility to introduce changes.

P04-15

On the possibility to use Set I of Advanced Progressive Matrices as a Short Form of Standard Progressive Matrices: A preliminary study

Chiesi, Francesca (Department of Psychology, University of Florence, Italy)
Ciancaleoni, Matteo (Department of Psychology, University of Florence, Italy)
Galli, Silvia (Department of Psychology, University of Florence, Italy)
Primi, Caterina (Department of Psychology, University of Florence, Italy)*

Abstract

A potential limitation of the original 60-item Standard Progressive Matrices (SPM) is test time administration that may be too long for some purposes. To reduce the length Nathaniel-James and colleagues (2004) suggested to use the 12 items of Set I of Advanced Progressive Matrices (APM) as a short form of the SPM. Indeed, APM–Set I items are characterized by an increasing difficulty level, and they cover the intellectual processes and the full range of difficulty sampled by the SPM. Since the psychometric properties of the APM–Set I were not investigated yet, the purpose of the present work was to preliminarily evaluate the possibility to use it as a short form of SPM applying Item Response Theory. The test was administered to a large sample of primary and middle school students (N= 998), consistently with the shared assumption that the SPM can be adequately used to assess general mental ability in children and adolescents from 8 to 16 years of age. Once the prerequisite of unidimensionality was confirmed, model fit analyses showed that three-parameter model (3PL) is the most suitable in line with previous results on the full-length SPM (Raven et al., 2005). Parameters attested that for the most part items difficulty levels (b) were consistent with the way in which items have been selected to increase in difficulty, items showed medium to large discrimination levels (a), and guessing (c) was found to be relevant for some items.

P04-17

Developing and using vertical scales in assessing growth

Rogers, H. Jane (University of Connecticut, USA)*
Swaminathan, Hariharan (University of Connecticut, USA)

Abstract

The issue of documenting the growth of children as they progress through the grades has attracted attention in recent years as a result of the movement to hold educators accountable for children's learning. The No Child Left Behind legislation (NCLB) in the USA provided the impetus for the accountability movement and forced states to demonstrate Adequate Yearly Progress (AYP) to show the effectiveness of instructional practices. The Status Model that compares the proportion of children who reach proficiency in one year with the proportion of children in the same grade who reached proficiency in the following year does not provide information about growth since different groups of children are involved. Assessing the effectiveness of instruction requires a measure of growth for each child. This requirement leads directly to a vertical scale that measures the growth of a child across grades. The development of a vertical scale requires considerable planning with respect to all aspects of testing. The purpose of this paper is to document the various aspects of the processes involved in developing and validating a vertical scale as implemented in a state in the USA. The paper describes and discusses test specifications, test design, data collection design, scale development, psychometric procedures for scaling, testing assumptions underlying the psychometric models, scale adjustments, standard setting, and reporting and interpreting of scores. In particular, the procedures developed by the authors for dealing with such psychometric issues as scaling/equating, scale adjustment at the tails, and establishing standards across grades are described.

P04-18

Cattell's T-data in the ICT society

Shih, Pei-Chun (Universidad Autonoma de Madrid, Spain)*
Martinez-Molina, Agustin (Universidad Autonoma de Madrid, Spain)
Montoro, Alejandra (Universidad Autonoma de Madrid, Spain)
Lopez-Almeida, Patricia I. (Universidad Autonoma de Madrid, Spain)

Abstract

The development of computer technology has enabled the design of performance tests to become easier than in R. B. Cattell's time. Moreover, the administration, scoring and accuracy have been improved, giving us more possibilities when assessing personality through T-data. The aim of this poster is to present the main results obtained in our research about objective assessment of personality in Risk Tendency, Trust Propensity and Conscientiousness. Results have shown good ranges of reliability and validity indices in six computerized behavioral tests of personality: internal consistency (i.e. Risk Propensity Dilemmas Task, α = .77, n = 892, Botella et al., 2008), temporal stability (i.e. Conscientiousness Tree Targets Task, 1-year test-retest: .51, n = 413, Hernández et al., 2003), concurrent validity (i.e. correlations between Tree Targets and Token Tasks: .64, n = 267, Hernández et al., 2009), and predictive validity (i.e. Risk Tendency and Risk Behavior, coefficient of determination = .55, n = 83, Shih et al., 2007). Therefore, following Cattell, we consider T-data computerized instruments as a very interesting psychological assessment method that will allow us to improve our scientific knowledge in personality psychology.

P04-19

Validity study on automatic scoring methods for the summarization of scientific articles

Su, I-Hsiang (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)*
Hung, Pi-Hsia (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)

Abstract

Summarization clearly has a great potential for improving students' reading, learning, and writing. Unfortunately, it is largely neglected throughout children's academic training due to the demanding performance assessment. In the current study, we create an automatic scoring system based on concept map-like Pathfinder network representations (PFNets; Shavelson & Ruiz-Primo, 2000) for text summarization of scientific articles written in Mandarin. The Pathfinder network analysis method is adopted in term of the knowledge structure evaluation. A prototype system integrated ALA-Reader and Pathfinder PCKNOT software (Schvaneveldt, 1990) are devised for converting the proximity data into visual PFNets from participants' text summaries. Two similarity indexes GTD (Graph-Theoretic Distance) and PFC (Pathfinder Correlation) (Goldsmith et al., 1991) are automatically derivate from pathfinder network analysis for scoring summarization. They are applied to evaluate knowledge structure underlying the summaries between a domain expert and four hundred and sixteen 5th and 6th graders in Southern Taiwan. The validity of scoring is discussed by comparing the correlation coefficient of human rater scores to the resulting PFNets agreement-with-expert scores (r = 0.63) with two commonly used alternatives, i.e. keyword pattern matching (r = 0.64) and LSA (Latent Semantic Analysis, r = 0.56). The result suggests the scoring method based on concept map-like representations is very promising to provide valuable feedbacks for learning.

P04-20

Application of a Hierarchical Bayesian Testlet model to TOEIC

Ra, Jongmin (The University of Georgia, USA)
Lee, Sunbok (The University of Georgia, USA)
Koh, Bora (Konyang University, Korea)
Wang, Aijun (The University of Georgia, USA)*

Abstract

Application of a Hierarchical Bayesian Testlet model to TOEIC Theoretical Background Standardized educational and psychological tests like the Test of English for International Communication (TOEIC) have been used to measure individuals' latent trait. Item response theory (IRT) offers mathematical models in which the relationship between the latent trait of interest and the test items can be specified and provides the probability of a correct response to an item in terms of item parameters (a, b, and c) and individual's latent trait. A testlet comprised of a set of items that share a common stimulus such as reading passage and essay may violate the assumption of local independence assumption since there is possibility of dependency among items within a testlet. The Bayesian testlet model suggested by Bradlow, Wainer, and Wang (1999) can incorporate the local dependence existed among items within a testlet and SCORIGHT 3.0 (Wang, Bradlow, & Wainer, 2004) makes users more accessible to the Bayesian testlet models. Research Questions First was to conduct a basic simulation study to investigate testlet effects. Second, an empirical study applied the standard three-parameter logistic (3PL) model and the three-parameter logistic testlet (3PLT) model to the TOEIC which has not been studied in the IRT framework. Data Analysis Procedures In order to investigate testlet effects within the 3PLT model, WinBUGS 1.4 was performed and later, analogous analyses were performed via BILOG-MG and SCORIGHT computer programs for the comparison for the comparisons. Results will be presented later

P04-21

To rectify the examinee ability overestimation and underestimation under GRM

Jian, Xiaozhu (South China Normal University, USA)*
Zhang, Minqiang (South China Normal University, USA)

Abstract

Under two-parameter and three-parameter Logistic models, Jian, Dai & Peng(2007), and Rulison & Loken(2009)had found that the ability of the examinee was underestimated when he made wrong response on items which were too easy relatively for the examinee, that is, sleeping phenomenon (Wright, 1977); and was overestimated when he made right response on items which were too difficult relatively, that is, guessing phenomenon. The author designed an polytomous test which consisted of 35 items. After the examinee finished the test, he was given one extra polytomous item which difficulty parameter varied from -4 to +4 in 21 different cases, and some of the items were too easy or difficult for the examinee. Under GRM, it were found: (1) The examinee was overestimated when he got full score on items which were too difficult relatively; (2) The examinee was underestimated when he got zero score on items which were too easy relatively; (3) The examinee was underestimated when he got middle score on too easy items, and was overestimated when he got middle score on too difficulty item. Under GRM, the author improved the method that Mislevy and Bock (1982) proposed, and down-weighted the responses to items that was too easy or difficult for the examinee relatively. After down-weighting the responses, the ability overestimation or underestimation was rectified when he made right responses to too difficulty items or wrong responses to too easy items. After C, ; parameter added into GRM, the ability overestimation or underestimation was also rectified.

P04-22

Practical consequences of model misfit in assessing academic growth

Zhao, Yue (Educational Testing Service, USA)*
Hambleton, Ronald (University of Massachusetts, Amherst, USA)

Abstract

The purpose of the study was to investigate the practical consequences of IRT model misfit in assessing achievement scores, academic growth, and passing rates in a typical state-wide assessment program. To the extent possible, this study was designed to reflect common equating practices, and the major variable which was manipulated in the study was model misfit. A typical state assessment with 50,000 students was chosen. Results indicated that the typical level of model misfit (found in practice) would result in 1,420 examinees or almost 3% of the examinees being placed into different performance categories when the ability growth was 0.25 (or .25 of a standard deviation of proficiency scores), and 1,085 examinees or 2.2% of the examinees being placed into different performance categories when the ability growth is 0.50 (or .50 of a standard deviation of proficiency scores). The differences in the performance classifications due to the choice of a non-fitting model are of major, practical consequence. The key messages from the study are that (1) practical ways are available to study model fit, and, (2) model fit or misfit can have practical consequences that should be considered when choosing an IRT model for a state testing program.

P05 13:30-15:00 Room105

P05-1

Comparison of Rasch model and one parameter logistic model on croatian national assessment data

Curkovic, Natalija (National Centre for External Evaluation of Education, Croatia)*
Sabic, Josip (National Centre for External Evaluation of Education, Croatia)
Buljan Culej, Jasminka (National Centre for External Evaluation of Education, Croatia)

Abstract

In the international project “Development of Instruments in Croatian National Assessment” Croatian National Centre for External Evaluation of Education (NCEEE) cooperated with Institute for Educational Measurement, Netherlands (CITO). In analysis of Mathematics data One Parameter Logistic Model (OPLM) was applied. OPLM is a model where difficulty parameters are estimated and discrimination indices are imputed as known constants, which is an extension of the Rasch model. The aim of this paper is to compare two different IRT models: Rasch model and OPLM in the terms of fitting statistics and ability distributions. The goodness-of-fit of an item set with the OPLM is investigated with five goodness-of-fit statistics. Results of 1229 gymnasium students on three Mathematics domains were used for the purpose of this research. OPLM software were used in the analysis of results (Verhelst, Glas & Verstralen, 2005) was used. Estimation through Conditional Maximum Likelihood method showed superiority of OPLM in compare to Rasch model for the data that were used. When using OPLM, it is possible to find a model that will have better fitting statistics than Rasch model. OPLM combines attractive mathematical properties of the Rasch model with the flexibility of the two-parameter model. By imputing different discrimination indices for different items, OPLM extends the applicability of the Rasch model.

P05-2

The 2 x 2 achievement goal framework and intrinsic motivation among Filipinos: A validation study

Dela Rosa, Elmer (Central Luzon State University, Philippines)*

Abstract

In Western studies, the utility of the 2x2 framework (Elliot & McGregor, 2001) was established from within a diverse sample (e.g., Cury, et.al., 2006; Witkow & Fugilni, 2007). However, there is not much evidence for the theoretical and psychometric validity of the 2x2 framework among Asian populations, where some research suggest problems with some of the basic conceptual categories underlying the model. This study explored the applicability of the four-factor achievement goal model (mastery-approach, mastery-avoidance, performance-approach, and performance-avoidance) among Filipino high school students, with particular emphasis on the item analysis, reliability and validity of the measure, and studied whether the framework is predictive of intrinsic motivation. Results from item analysis provided a relatively strong support that the 2x2 achievement goal measure was internally consistent. Exploratory factor analysis showed only three distinct factors as against the hypothesized four-factor model. All items representing avoidance goals (mastery-avoidance and performance-avoidance) significantly loaded on a single factor. Confirmatory factor analysis was performed using both absolute and incremental fit indices. Results indicated that the data did not fit the model under investigation. Nevertheless, all of the achievement goals but performance-avoidance goals were significant predictors of students' intrinsic motivation.

P05-3

Validating the factors of the English and Filipino versions of the sense of self scale

Ganotice, Fraide (Palawan State University, Philippines)*
Bernardo, Allan Benedict (De La Salle University, Manila, Philippines)

Abstract

The Sense of Self (SOS) Scale measures different aspects of a learner's sense of purpose and self-concept that are related to their motivations to strive in their academic endeavors, and is designed to have four dimensions: sense of purpose, sense of reliance, negative self-esteem, and positive self-esteem. In this study, we developed a (conversational) Filipino version of the SOS using a translation process that combined a committee approach with back-translation. To assess the validity of the four-factor structure of the Filipino and English versions of the SOS, 765 high school students were asked to complete one of the two versions. Confirmatory factory analysis indicated a good fit between the four-factor structure and the data from the two language versions. The scales corresponding to the four factors were also found to show adequate internal reliability for both versions. Finally, the pattern of correlations among the four factors was similar for both versions. The discussion focused on the viability of the four-factor structure of the Filipino and English SOS Scales for use in various research explorations on Filipino students' motivations in school.

P05-4

The measurement of off-line meta-cognitive regulation and its application on personnel selection

Li, Jian (School of Psychology, Beijing Normal University, Beijing, China)*

Abstract

The purpose of the present study was to explore the psychometric structure of off-line meta-cognitive regulation, and to construct assessing instrument with high criterion-related validity. The data of this study were collected from a large-scale IT company in China. In study 1, 260 employees participated in the preliminary test. By using a self-developed questionnaire, a three-factor structure of off-line meta-cognitive regulation was obtained, which includes global-planning, insight and generalization. After a retest study with 115 participants, the structure of off-line meta-cognitive regulation was confirmed. In study 2, in order to examine the criterion-related validity of the off-line regulation questionnaire, hierarchical linear regression and hierarchical logistic regression were conducted by the researcher. The following findings were obtained: The managers are different from the common staffs in particular regulation factors. Managers have higher generalizing level than others in sales departments, while managers in support department have higher insight level than others. But on the other hand, for the common staffs, the level of off-line meta-cognitive regulation fails to predict employees' performance in sales department. However, for the staffs in support department, the level of global-planning predicts the creative activities reversely.

P05-5

Cognitive analysis on item difficulty of Matrix Reasoning Test

Li, Zhongquan (Department of Psychology, School of Social and Behavior Sciences, Nanjing University, China)*

Abstract

Computerized automatic item generation (AIG) was gradually recognized as a promising technique in dealing with item exposure. Understanding sources of item variation was the initial stage for Computerized AIG. The present study investigated the relation between matrix reasoning item difficulties and stimuli factors such as familiarity of figures, abstract of attributes, perceptual organization, the numbers of elements as well as memory load. 8 sets of matrix reasoning tests were constructed by manipulating those factors. Using anchor-test design, these tests were administrated via internet among 1929 participants with 10 APM items as anchor items. After unidimensional and local independence established, those items were calibrated with BILOG-MG 3.0 (Marginal maximum likelihood estimation and 2PL model), and proved to pose good item difficulties and discriminations. Item parameters were equated using IRTEQ to make items from different sets comparable. ANOVA and regression analysis indicated that all but familiarity of figures could significantly predict item difficult, and memory load (i.e. combination of types and number of rules) was the most important predictor. The findings were important for Computerized AIG, since new items with predicted item difficulties could be generated by manipulating those factors.

P05-6

Psychometric properties of Classmates' Friendship Questionnaire

Turilova-Miščenko, Tatjana (University of Latvia, Department of Psychology, Latvia)*
Raščevska, Malgožata (University of Latvia, Department of Psychology, Latvia)

Abstract

The purpose of this research was to determine the psychometric properties of the Classmates' Friendship Questionnaire (CFQ). The sample consisted of 264 participants from Latvian schools aged from 13 to 15 years (boys – 40 %, girls – 60 %). The Peer-Relations subscale (PRs) of the Self-Esteem Questionnaire (Hunter, Boyle, & Warden, 2006) was used in order to verify the concurrent and convergent validities of the CFQ. The factorial validity of the CFQ was established using principal components analysis with varimax rotation; this yielded four factors: Trust, Support and Collaboration, Social Contacts out of School, and Lack of Hostility. All CFQ scales had high internal consistency and test-retest reliability, and correlated significantly with the PRs.

P05-7

Performance of urban and rural adult populations on neuropsychological tests in Zambia

Zulu, Happy (The University of Zambia, Zambia)*

Abstract

Introduction Considering the high prevalence of neurobehavioural deficits due to illnesses such as malaria and HIV/AIDS in Zambia like in many other Sub-Saharan Africa countries, effective and efficient assessment of neurobehavioural function is significant. Objectives This study is mainly aimed at examining the performance of the adult Zambian population on neuropsychological tests with regard to the urban-rural dichotomy and formulation of the normative data for Zambia. Methods The research will be carried out on 324 HIV negative participants; urban (n=162) and rural (n=162); age 18 to 65years; and with primary to tertiary education. Their scores on the Zambia Neurobehavioural Test Battery assessing domains like verbal fluency, working memory and speed of information processing. ANOVA will be used to analyze the data. Results Discrepancy in performance on neuropsychological tests between urban and rural areas of Zambia will be determined and normative data formulation for Zambia. Conclusion It will enhance appropriate utilization of neuropsychological tests in the local setting against the background of issues related to psychological tests in general such as cultural bias.

P05-8

The types and nature of questions vis-à-vis students' test-taking skills as significant indicators of second language examinees' performance on the TOEFL-ITP reading comprehension sub-test

Amurao, Analiza Liezl (Mahidol University International College, Thailand)*

Abstract

This study examines the reading performance of selected students at the Pre-College program of the Mahidol University International College (PC-MUIC) as they are required to attain a score of 520 in the TOEFL-ITP (or equivalent performance in IELTS) to enter MUIC. Specifically, this research aims to evaluate whether the reading skills that examinees possess correlate with successful performance on the Reading Comprehension sub-test of the TOEFL-ITP. Only TOEFL-ITP Reading Comprehension Sub-test performance has been considered in this study as IELTS is not taught or administered in the Pre-College program. This study makes use of descriptive qualitative design relying heavily on the following instruments for data collection: Commercial-based test-prep texts (Reading Comprehension Sub-section), Schraw and Roedel's Levels of Difficulty (1992), the researcher's modification of said band, the respondents' scores per question type, tabulations of the respondents' scores based on the levels of difficulty of the items and the question types used in the test, focused interviews with the respondents, and retrospective journal entries of the researcher. This study aims to shed light on issues surrounding how second language learners' reading skills affect performance on standardized tests such as TOEFL. This study specifically seeks to provide MUIC PC instructors empirical data that would help them understand their own students' reading difficulties which, consequentially, will aid them address teaching-learning issues.

P05-9

Dynamic testing: A missing element on the assessment "of learning" and assessment "for learning dichotomy"

Arce-Ferrer, Alvaro (Pearson, USA)*

Abstract

This study investigates the incremental validity of dynamic measures over and above static measures of assessment "of learning" and assessment "for learning" in the context of problem-solving skills within introductory high-school physics. A typology of assessment gaining impetus in multiple nations typifies assessments into two mutually exclusive categories: assessments "of learning" and assessments "for learning." Whereas assessments "of learning" and assessments "for learning" purposefully controls for instruction when assessment takes place, dynamic assessments provide measures of examinees' ability to profit from instruction not available under their static administrations. Dynamic testing combines unassisted testing (i.e., static testing) with explicit instructional and feedback scaffolding crafted within the test to assist examinees performance during the execution of a purposefully difficult task. A training-within-the test computerized assessment system was developed and loaded with physics problems and instructional prompts hierarchically organized based on problem solving principles derived from research and classroom practice. A sample of high school students, novice in physics problem solving, answered unassisted pretest measures of assessment "for learning" (i.e., inductive reasoning, reading comprehension) and assessment "of learning" (i.e., high school physics test). During the training phase, dynamic measures on the physics problems were collected for each examinee. In a posterior session, participants' unassisted post-test measures of assessment "of learning" with far and very-far-transfer tasks were collected. Preliminary results showed the incremental predictive validity of dynamic measures over and beyond static measures of assessment "of learning" and assessment "for learning."

P05-10

A new strategy of item selection of cognitive diagnosis computerized adaptive testing

Du, Xuan-xuan (Jiangxi Normal University, China)
Ding, Shu-Liang (Jiangxi Normal University, China)*
Gan, Deng-wen (Jiangxi Normal University, China)

Abstract

As an important improvement in the test history, Computerized Adaptive Testing (CAT) has unparalleled advantages. With the development of Cognitive Diagnosis (CD) which can provide diagnostic information about students' misconceptions, the Cognitive Diagnosis Computerized Adaptive Testing (CD-CAT) has become an irreversible trend in modern educational assessment. The most commonly item-selection strategies of CD-CAT are MI (Maximum Information method) (Lord,1977),KL(Kullback-Leibler Information Method) (Chang & Ying,1996;Xueli Xu,2003) and SHE(Shannon's Entropy Method) (Tatsuoka,2002;Vomlel 2003;Xueli Xu,2003). Compared in these three item-selection strategies, MI has an advantage on accurate and efficient of the result of estimate Ability of examinees but inefficient at the result of estimating examinees' Knowledge States. On the contrary, the result of

SHE is the most accurate and efficient at estimating of Knowledge State but inefficient at estimating of Ability. And as compared with SHE, KL has no advantage on neither Ability nor Knowledge State. Thus, a new method was found for providing more accurate and efficient estimate of examinees' both Ability and Knowledge State at the same time. A new combined strategy was discussed in this study. The combined strategy is that the selected item maximizes the Fisher information at current value of ability from the candidate items which minimize the Shannon entropy from the Item Bank. The administered item is optimal for accuracy of Ability estimation and Knowledge State estimation. The new strategy not only saves the time but also saves the financial costs, also balances the item exposure.

P05-11

Comparing item performance between domestic and international examinees on a high stakes licensure computerized adaptive test

Gorham, Jerry (Pearson, Inc., USA)*

Woo, Ada (National Council of the State Boards of Nursing, USA)

Abstract

For many practical reasons, international examinations are often developed and scaled in the country of origin and subsequently expanded using test and item characteristics derived from the domestic samples. Assumptions about the comparability between item characteristics based on the domestic population vs. the extended populations should be tested, if possible, to understand the robustness of the exams and to ensure validity and precision of scores. This study will investigate whether item statistics based on US samples provide a reasonable application basis for international examinees for a high stakes licensure computerized adaptive testing program. We will compare data collected from US and international samples on a high stakes licensure CAT to evaluate the degree of consistency between domestic and international performance on the test. Samples will be collected from the top 10 highest volume countries and compared to the US population samples. Our method will focus on differential item performance comparisons to detect potential group differences on items. We will also investigate factors such as English Second Language indicators as a possible explanation of any potential differences. Overall test properties, distributions of ability estimates, item parameter estimates, and ancillary data elements such as item response time will also be included in the analysis. Preliminary results indicate the need for regular monitoring of exam data to ensure against item compromise or differential item performance in international test sites compared to domestic populations.

P05-12

Prediction of students' academic performance based on bootstrap method and ARIMA model

Han, Yuna (South China Normal University, China)*

Zhang, Minqiang (South China Normal University, China)

Abstract

Making scientific prediction of the students' academic performance can allow both the students and the teachers to have a clear mind about their own destinations and ways to work hard. It also provides the basis for future teaching and learning. Traditionally, we have averaging method, regression analysis and other methods for predicting students' academic performance. In this survey, we use the mathematic weekly test results of a group of students in Grade Twelve as the sample to elaborate how to use the Bootstrap method and the ARIMA model to make prediction of students' academic performance. Meanwhile, by using the same sample, we compare such method with the traditional methods to check out the precision. As the result, we find that by applying the ARIMA model we could get the more accurate estimated value than the traditional methods. Moreover, using the ARIMA model based on Bootstrap method provides the better perdition than any others.

P05-14

The latent class analysis of creativity performances

Lee, Pei-Yu (National Taiwan Normal University, Taiwan)*

Chen, Po-Hsi (National Taiwan Normal University, Taiwan)

Abstract

The purpose of the research is to apply the latent class analysis on creativity performances and investigate the relationship between different types of creativity performances and their creating process. Two hundreds and seventeen university students were asked to answer a creativity design test. In this test, subjects should choose 3 geons from 36 geons, change and assemble them into some kinds of furniture. Their creating process and product were both scored using seven criterions by three of two raters. Data analysis was separated into two stages. In the first stage, the latent class analysis was used to analyzed the seven product scores and sort these subjects into different types of creativity performances. Results showed that three categories fit better than other number of categories. These three groups of subjects were named as "Low Innovative - High Practical", "High Innovative - High Practical", and "High Innovative - Low Practical". In the second stage, analysis of variance (ANOVA) was used to compare the creating process scores of different groups of subjects. Results showed that High Innovative group is significantly different from Low Innovative group in some creating process scores. The High Practical group is also significantly different from Low Practical group in some creating process scores. These results indicated that creating processes are related to the innovation and practicality of the creative products.

P05-15

Integration of augmented reality into the teaching of spatial concepts

Lin, Chien-Yu (Graduate Institute of Assistive Technology, National University of Tainan, Taiwan)*

Hung, Pi-Hsia (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)

Abstract

This study used augmented reality technology to combined real scenery and three dimension virtual. When detected by the web-cam device, the corresponding information was shown on screen to improve the sense of three-dimension and the interaction of the spatial assessment. The main purpose of this study is to apply augmented reality to assessment

for spatial learning units. The key components are identifying tags for reality augmentation, web-cam devices and image processing devices. By applying this technology, it is possible to see stacked objects through all 360 degrees. The resultant visual presentation is very different from what can be achieved with assessment on paper or in videos. Through the interaction of the users and the three dimension virtual objects in this study, this design of teaching materials offered both teachers and students a novel learning method. The use of augmented reality in this study for this area offers a new direction of development for assessment applications.

P05-16

Impact of computer technologies upon cognitive abilities of students

Cheremoshkina, Lubov (Moscow State Pedagogical University, Moscow, Russia)
Molchanov, Alexander (Moscow State University of Medicine and Dentistry)*

Abstract

This experimental investigation concerns cognitive activity of software programmers, PC users with different experience, Internet users, and PC gamers. Various activities of PC users impact on theirs cognitive ability in a different way. 1. PC and Internet users with more than three years experience have significantly higher cognitive and mnemonic abilities, than users with the experience from one till three years. 2. Internet users have higher efficiency of cognitive and mnemonic abilities, than PC users. 3. Software programmers have higher efficiency of cognitive and mnemonic abilities, than Internet users. 4. Prolonged experience of computer technologies using increase the analyticity of perceptive processes. 5. Prolonged work in Internet decrease visual memory efficiency for the reason of obvious mnemonic abilities control mechanisms deformation. 6. Active PC gamers differ from the contemporaries by decrease of mnemonic abilities efficiency and by increase of reaction time in response to visual, tactile and acoustical stimuli, during both relaxation, and intellectual loading. 7. The neuropsychological analysis of PC gamers mnemonic processes has allowed to identify signs adverse for effectiveness of memory: low activation of nervous system; inadequate hemispheric activation to displayed stimuli; level of PC gamers' nervous system activation is inadequate to complexity of intellectual loading. Remembering of simple information evokes sympathetic hyperactivation.

P05-17

Investigating item drift in TIMSS via Cognitive Diagnostic Modeling

Park, Yoon Soo (Teachers College, Columbia University, USA)*
Lee, Young-Sun (Teachers College, Columbia University, USA)

Abstract

Longitudinal studies of trended international mathematics achievement have focused on the overall performance of students using trended assessments such as the Trends in Mathematics and Science Study (TIMSS) and the OECD Program for International Student Assessment (PISA). They have employed methods such as Classical Test Theory (CTT) and Item Response Theory (IRT) to rank individuals within a latent ability continuum. Although inferences generated from these approaches have provided insights into the relative standing of students and their mathematics ability in comparison to other countries, they have yet to examine how specific attribute mastery change over time—whether unique skills required to solve a mathematics problem have grown or remained constant with time. This view is different from examining student performance in broad domains such as algebra and geometry, because investigating fine-grained attributes form the basis to students' understanding of the material as well as providing direct information to educational researchers and instructors on areas that students need improvement. Cognitive Diagnostic Models (CDM) were developed for this specific purpose—to examine whether a specific cohort has mastered skills that are required to correctly answer a problem. Using three waves of the TIMSS—1999, 2003, and 2007—this study examined attribute mastery from a CDM framework while implementing longitudinal analysis methods. This study shows that by examining item drift—whether CDM parameter estimates change over time—we can signal instructors and mathematics researchers on content areas that have deviated and detect trends in attribute mastery that cannot be estimated using traditional methods.

P05-18

A study on differential item functioning (DIF) of the Basic Mathematical Competence Test for junior high schools in Taiwan

Cheng, Chien-Ming (National Academic for Educational Research Preparatory Office, Taiwan)*
Wang, Hsuan-Po (National Taichung University, Taiwan)

Abstract

This study investigates the relationship between a gender's group membership and performance on test items using four differential item functioning procedures – Area Measure, Likelihood ratio test, Mantel-Haenszel, and SIBTEST methods. The basic competence test for junior high schools is a very important breakthrough in education in Taiwan because it adopts the item response theory (IRT). The DIF topic in IRT is important because of concern that the basic competence test for junior high schools be fair and impartial for every student. In the study the presence of DIF for gender groups is investigated for this new system of testing. The results of this study are the identification of items that show evidence of DIF and that are judged to be due to bias, a determination of which methods are the most accurate for detecting DIF and an investigation of the possible reasons for causes of DIF and bias. Both real and simulation data are analyzed to compare the four detecting DIF methods. From the results, synthesis and discussion of effect size, frequency, consistency, and Type I error rate, of the four methods, SIBTEST was deemed the most appropriate to detect DIF items for the basic mathematical competence test for junior high schools in Taiwan.

P05-19

Applying the many-facet Rasch model to evaluate concept map for statistics in graduate education

Wang, Li jun (College of Education, Zhejiang Normal University, China)*
Wang, Wen Chung (The Hong Kong Institute of Education, Hong Kong SAR, China)
Gu, Hai Gen (Shanghai Normal University, China)

Abstract

Concept maps are a technique for organizing knowledge graphically. Concept maps give the possibility to capture, process, and store information, in a way similar to that used by the brain. Concept maps carry out a schematic representation of significant relationships among concepts expressed as sentences. When the apprentice is able to relate the new concepts to previous knowledge within his/her own cognitive structure, learning takes place. There are authors in statistics that already are using concept maps in their work. Concept maps are helpful as a tool to gauge students' understanding because they make the knowledge construction process visible. This study investigated the usefulness of the many-facet Rasch model (MFRM) in evaluating the quality of performance related to concept map for statistics in graduate education. 21 Chinese graduate students prepared concept map for statistics after a half-year course. The students choose some topic in statistics, based on the content and objectives of the course. Eight Statistics Specialists evaluated each student's concept map, based on three components of concept maps, including knowledge amount, theory level, applying ability. Three components were scored on a five-point scale (1 = none, 5 = excellent). The results of this study show that the MFRM technique is a powerful tool for handling polytomous data in performance and peer assessment in graduate education.

P05-20

Are negatively worded items doing a good or bad job? A study of wording effects in measuring self-esteem

Cheng, Christopher H K (City University of Hong Kong, Hong Kong SAR, China)
Ye, Shengquan (City University of Hong Kong, Hong Kong SAR, China)*

Abstract

It is quite common for self-report instruments to include both positively and negatively worded items so as to reduce the so-called response set bias such as acquiescence or agreement bias. However, it has been pointed out that this strategy (including both positively and negatively worded items) could complicate the underlying measurement model by adding a spurious latent factor to the original content dimension. The present study is to evaluate the effects of including both the positive and negative items in an emically developed self-esteem instrument, the Chinese Adolescent Self-Esteem Scales (CASES; Cheng, 2005). The General Self subscale (GS) of the CASES consists of 8 items, half of which are negatively worded. A series of models were evaluated: single trait (self-esteem) factor model (M1), two factors (negative esteem and positive esteem) model (M2), one self-esteem factor and one "positive" method factor (M3), one self-esteem factor and one "negative" method factor (M4), and one self-esteem factor and both "positive" and "negative" method factors (M5). Although the unidimensional model (M1) was not rejected, it had significantly poorer fit than the two-factor model (M2). Except M1, all models had very high and similar fit but findings were not conclusive. While M3 and M4 provided support for the method effect of positive wording and negative wording respectively, the factor loadings of "negative" method factor in M5 were not statistically significant, thus challenging the existing literature about the unfavorable effects of negative wording. More evidence is needed in future research before the issue can be clarified.

P05-21

Application of neural networks in the creativity measurement for middle school students

Yu, Jiayuan (Nanjing Normal University, China)*
Wu, Rongping (BenQ Corporation, China)

Abstract

Multiple regression has frequently been used in psychological test for predict some dependent variable from a group of independent variables. One major challenge, however, is that all these variables have to satisfy the requirement for normal distribution and linear relationship between variables. Otherwise, neural networks could be applied. In this study, the scientific creativity was predicted by creative personality score. Neural networks and linear regression models were set up respectively. The predictive accuracy of these models was compared. 550 students of middle school in two cities were measured with Williams Creativity Assessment Questionnaire (WCAQ) and Adolescence Creativity Scale (ACS). The data of 70% students was used to train neural networks and build regression model, other 30% data was used for testing and prediction. Scores of four dimensions in WCAQ were used as input of neural networks and independent variables of regression model respectively. Scores of ACS were used as output of neural networks and dependent variables of regression model respectively. The results shows the neural networks could predict students' scientific creativity more accuracy.

P05-23

Determining the factors of students' parents actions towards academic achievement using structural equation modeling

Koh, William (University Science Malaysia, Malaysia)
Ong, Saw Lan (University Science Malaysia, Malaysia)*

Abstract

Parental influence is a vital aspect in students' academic achievement and discipline. Various studies overseas have found that what parents do have a significant relationship with their children school's outcome. Hence, this dictate the urgency of such study to be carry out in Malaysia. However, the existence of an instrument to determine the parental involvement constructs is scarcely available. Therefore, the main objective of this study is to develop and validate "Student's Parents-Actions Questionnaire version 1" (SPAQ1). SPAQ1 is an instrument developed specifically to determine the various dimensionality of parental involvement with students in the Malaysian context. A few theories such as Hirschi's Social Bonding Theory, Attachment Theory (Bowlby, 1969), Social Learning Theory (1976) and Social Cognitive Theory (Bandura,1986), Overlapping Spheres of Influence (Epstein, 1997) and the Ecological System Theory (Bronfenbrenner,1979), were used as the underpinning theories in the item formulation development process. Steps involved include coining and formulating related items, defining the variables operationally and validation of the instrument. The early version SPAQ1 consists of 62 Likert-scale items (α = 0.96)and was refined and reduced to 45. The content validity, construct validity, and factorial validity across ethnics and reliability of the instrument were determined. Configural invariance across the major ethnics in Malaysia namely the Malays, Chinese and Indians is assessed using Structural Equation Model. The measurement model for the instrument was constructed and it shows convergent and discriminant validity. The availability of SPAQ1 will contribute to the drawing up of inter-ethnics parental training programmes in Malaysia.

P06 15:15-16:45 Room 105

P06-1

Reliability and factorial validity of Farsi version of the Positive and Negative Perfectionism Scale (FPANPS)

Besharat, Mohammad Ali (University of Tehran, Iran)*

Abstract

Perfectionism is a personality trait characterized by striving for flawlessness and setting excessively high standards for performance, accompanied by tendencies toward overly critical evaluations of one's behaviour. Research has suggested that two major dimensions of perfectionism, positive and negative perfectionism, should be differentiated. This study investigated reliability and factorial validity of a Farsi version of the Positive and Negative Perfectionism Scale (PANPS; Terry-Short, Owens, Slade, & Dewey, 1995) for 606 undergraduate students (257 males and 349 females) from the University of Tehran. All participants were asked to complete the Farsi version of the Positive and Negative Perfectionism Scale (FPANPS; Besharat, 2005), the General Health Questionnaire (GHQ; Goldberg, 1972) and the Coopersmith Self-Esteem Inventory (SEI; Coopersmith, 1967). Findings supported the internal consistency, test-retest reliability, concurrent validity, and two-factor structure of the FPANPS. The factors found in the FPANPS are similar to the factors found in previous studies and were accordingly labeled as Positive Perfectionism and Negative Perfectionism. The results provide evidence for applicability of the FPANPS for Iranian populations and its cross-cultural validity.

P06-2

Temperament styles of children from Taiwan, the PRC, and South Korea

Chang, Te-Sheng (Graduate Institute of Multicultural Education, National Dong Hua University, Taiwan)*
Chen, Hsin-Yi (Department of Special Education, National Taiwan Normal University, Taiwan)
Oakland, Thomas (Department of Educational Psychology, University of Florida, USA)

Abstract

By applying the Student Styles Questionnaire(SSQ; Oakland et al., 1996), gender, age, and cross-national differences of children ages 9 through 17 in Taiwan, the People's Republic of China, and South Korea are examined on four temperament styles: extroversion-introversion, practical-imaginative, thinking-feeling, and organized-flexible styles. In general, results based on 1929 children from Taiwan revealed that, Taiwan children prefer practical to imaginative styles, feeling to thinking styles, and organized to flexible styles. Their preferences for extroversion- introversion styles are more balanced. Gender differences are found on practical-imaginative, thinking-feeling, and organized-flexible styles. Age differences are found on extroversion-introversion, thinking-feeling, and organized-flexible styles. In contrast to children in Taiwan, those in the PRC are more likely to favor extroverted, practical, thinking, and organized styles while those in South Korea are more likely to favor extroverted, imaginative, feeling, and flexible styles. Gender differences appear with all three samples: girls are more likely to favor a feeling style and boys to favor a thinking style. Detailed discussions regarding the temperament styles of Asian children and children from other nations are provided. Implications of this study are also discussed.

P06-3

The dark side of courtship: Traditional behaviors in dates and sexual coercion in a university sample

García Meraz, Melissa (Universidad Autónoma del Estado de Hidalgo/ Autonomous University of Hidalgo State, Mexico)*
Del Castillo Arreola, Arturo (Universidad Autónoma del Estado de Hidalgo/ Autonomous University of Hidalgo State, Mexico)
Martínez Martínez, Juan Patricio (Universidad Autónoma del Estado de Hidalgo/ Autonomous University of Hidalgo State, Mexico)

Abstract

Recent research indicates that sexual coercion can be experienced by females and males in similar rates. It also has been related to traditional behaviors in dates, for example: males assume the key roles in dating, including initiating the date, absorbing the monetary cost, taking the protector role and calling the girl after the date asking if she spend a good time. In the other side, females assume the contrary role, controlling the sexual advances of males and given the maintenance behaviors in the relation. In order to explore behavior related to sexual coercion and traditional attitudes, 90 in-depth interviews were conducted in both females and males. With the results two scales were develop and applied to 300 undergraduated students. Results show that males and females use sexual coercion in their relationships, they use rational arguments, lie, remember their duty as man or woman, make promises that are not going to fulfill and promise eternal love. When males and females assume traditional behavior, they increase the use of sexual coercion. This results show that sexual coercion is an important and yet under-recognized topic, practiced by females and males, even when is more practiced by males.

P06-4

The problem of measurement invariance in Deci and Ryan's self-determination theory

Hanfstingl, Barbara (University of Klagenfurt, Austria)*
Gnambs, Timo (University of Linz, Austria)
Andreitz, Irina (University of Klagenfurt, Austria)
Thomas, Almut (University of Klagenfurt, Austria)
Florian H. Mueller (University of Klagenfurt, Austria)

Abstract

Deci and Ryan (e.g. 2002) differentiate intrinsic and extrinsic motivation in a more precise way. They postulate four different regulatory styles that can be distinguished empirically. The aim of the presented paper is to show that measurement invariance is not given to the construct of motivation. Basis of the results is a large-scale study with 1992 students in grades 5 to 9 and their teachers (N=89) taking part at the so called IMST-Project. This project supports teachers who test innovative instructional styles in schools. The authors discuss possible causes of the not given measurement invariance.

P06-5

The assessment of behavioral disorders in children who are deprived of parents.

Hatami, Mohammad (Tarbiat Moellem University, Iran)*

Abstract

The aims of this study are to assess behavioral disorders in children who have lost their parents in the earthquake. In order to do this, selected 32children (boys and girls) between 6-11 years old that live in boarding –houses & family houses and using research made questionnaire and Rutter questionnaire, interview and observation, according to: age, sex, birthrate, age of losing parents and stay in the Boarding-house and family house, were evaluated. With the descriptive method results showed that: 1- Sixteen percent of the subjects completely, thirty two percent to some extent had all types of disorders but in fifty two percent no disorders have been observed. 2-Attention deficit disorder has shown the greatest prevalence and the antisocial behavior disorder has shown the least of prevalence. 3-different types of disorders have been observed mostly in girl rather than boys, in younger children more. 4- Stay in the Boarding-house has caused more behavioral disorders.

P06-6

To develop of natural science academic aptitude tests for high school students in Taiwan

Hou, Ya-Ling (Department of Special Education, National Pingtung University of Education, Taiwan)*

Abstract

Scientific talent is one of important categories for gifted and/or talented students. There are many scientific talented classrooms provided for high school gifted and /or talented students in Taiwan. Standardized Aptitude Test can be an excellent indicator to identify scientific talented students' potentials. The main purpose of this study was to develop two types (Test A and Test B) of National Science Aptitude Test (NSAT) in order to identify whether the students have the scientific talented potentials. The researcher first implemented test equating in order to establish a common scale by three parameters (discrimination, difficulty and guessing parameters). Then, Item Response Theory (IRT) applied to execute test editing. In addition, the target information function (the researcher used +2) for gifted/talented test hoped to provide maximum discrimination and information for students at ability of +2. Both types of the NSAT had 45 items including Physics, Chemistry, Biology and Earth Science knowledge. Coefficients of stability for two weeks achieved .854 and .887. Coefficient of stability and equivalence was .799. The criterion-related validity evidence from estimate of the correlation between "The Basic Competence Test" scores and the NSAT scores were .60. It was indicted that there was not significant gender bias for all items by the gender Differential Item Functioning. The researcher also established the percentile rank and normal T-score norm based on 1358 high school students in Test A and 1224 students in Test B.

P06-7

Measurement invariance of the resilience scale for Chinese adolescents across grade and gender

Huang, Duan (Institute of Psychology, Chinese Academy of Sciences (CAS), China)*
Zhang, Jianxin (Institute of Psychology, Chinese Academy of Sciences (CAS), China)

Abstract

The resilience scale for Chinese adolescents is a measure which measures 5 facets of resilience: goal planning, help seeking, family support, affect control, and positive thinking. Before the measure can be used to make gender and grade comparisons, it must exhibit adequate cross-group measurement invariance. Thus, the present study was designed to examine the grade and gender cross-group measurement invariance of the resilience scale for Chinese adolescents. A sample of 1582 high school students (770 boys and 812 girls) was recruited for this study, with 929 students from grade 8 and 653 students from grade 11.Multiple-groups confirmatory factor analysis were conducted to assess the measurement invariance of the scale, which began with the "full model" that tested the two grade groups by two gender groups (i.e., female students in grade 8, male students in grade 8, female students in grade 11, male students in grade 11). If the full model was not found to be invariant, then four grade and gender subgroups were tested (i.e., female model, male model, grade 8 model, grade 11 model).All groups met requirements for configural and metric invariance in the "full model". Scalar invariance and partial scalar invariance was found only for the grade 8 model and the grade 11 model, which indicated that mean comparisons may be conducted across genders for students in grade 8 and grade 11 but should not be conducted across grade groups within either gender.

P06-8

DIF analysis of the Creative Self-efficacy Scale in college students

Hung, Su-Pin (National Taiwan Normal University, Taiwan)*

Abstract

Since self-efficacy beliefs are domain specific, the scale of creative self-efficacy (CSE) is developed. However, the social climate in eastern culture may influence different attitudes toward to express creative self-efficacy in terms of gender. Thus, the present study aims to assess if any items in the CSE-student scale favors specific gender group by detecting of differential item functioning (DIF). Because of pure common metric is essential to DIF detection. Therefore, to preventing the common metric is contaminated by the inclusion of DIF items. The study adopts the constant-item (CI) methods for its superiority than other method has been proven in past simulation studies. Thus, the iterative constant procedure is implemented in this study to locate four DIF-free items as anchors in the CI procedure. A total of 636 undergraduates (270 males and 366 females) were recruited from six universities in Taiwan and the rating scale model (RSM; Andrich, 1978) was used to estimate item parameters. The mean square error (MNSQ) was adopted as model-data fit index. The main results: (1) all items of the CSE-student scale fit Rasch model appropriately. (2)Besides, the result of DIF detection with CI method found that the magnitudes of DIF among all items are less than .4, the mean magnitudes of DIF is 0.026 show that the scale does not exhibit gender DIF. (3) Finally, the mean latent trait of female students was lower than that of the male students on CSE-student scale. Suggestions and implications for future studies are addressed.

P06-9

The reliability and validity of the Goal Orientation and Learning Strategies Survey: A Filipino investigation

King, Ronnel (The University of Hong Kong, Hong Kong SAR, China)*
Watkins, David (The University of Hong Kong, Hong Kong SAR, China)

Abstract

The Goal Orientation and Learning Strategies Survey (GOALS-S; Dowson & McInerney, 2004) is an instrument designed to assess the academic goals, social goals, cognitive strategies, and meta-cognitive strategies of secondary school students. This instrument was initially developed and validated in Australia. The applicability of this instrument to the Philippine setting was tested in a study involving 1,000 Filipino secondary school students. Responses to this questionnaire are shown to have good internal consistency reliability and support is provided for its construct validity in terms of its factorial structure and correlations with other educational outcomes. Suggestions for further research using the GOALS-S are provided.

P06-10

The development of an online reading comprehension assessment

Chang, Kuei-Lin (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)*
Hung, Pi-Hsia (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)
Lin, Yin-Chien (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)
Chou, Lan-Fang (Graduate Institute of Measurement and Statistics, National University of Tainan, Taiwan)

Abstract

The nature of reading has been rapidly changing in the workplace as economic units seek to meet global economic competition. These days, information and communication technologies are a powerful driver for economy-wide productivity, growth and jobs. Online reading is not isomorphic with the offline reading. The IRA has recognized the unique nature of online reading comprehension; however, there is few empirical research about the online reading construct (Leu et al., 2007). The purpose of this study is to develop an assessment which combines cognitive and metacognitive aspects to investigate the construct validity of online reading. Research issues focus on discussing how the ICTs alter the nature of reading comprehension. The relationship between students' online reading comprehension and their online metacognitive strategy is also analyzed. There are two components for online reading: text processing and navigation. Three online reading units are developed in this study. There are 4 metacognitive items for each unit to assess the awareness about the usefulness of different reading strategies. The reading strategies included are locating information, analyzing information, synthesizing information, and communicating information. Two hundred 9th graders of southern Taiwan are sampled to take the online reading assessment. A standardized computer-deliver reading comprehension test and students' school grades are used for preliminary construct validity discussions. Online reading comprehension is reasonably independent from the standardized computer-deliver reading comprehension and students' school grades. The tasks of online searching and integrating large amount of information are functioning differently from the conventional reading tasks.

P06-11

Understanding academic engagement of Chinese university students: Confirmatory factor and latent class analysis

Li, Xueyan (Faculty of Education, The University of Hong Kong, Hong Kong SAR, China)*

Abstract

During the past twenty years, the concept of students' engagement has received increasing attention in educational psychology. In this study, Martin's Motivation and Engagement Scale-University/College (MES-UC), was adapted and administered to 832 Chinese university students. The cross-cultural validity of the scale was examined based on content analysis, reliability analyses, Confirmatory Factor Analysis, and Multiple-Indicator-Multiple-Cause modeling. The results show acceptable reliability of MES-UC in Chinese university student sample. The factor structure pattern of MES-UC identified by CFA matches that obtained by Martin's study (2005). Girls broadly reflect more adaptive motivation and engagement than boys; higher grade students reflect less adaptive motivation and engagement than their lower grade counterparts; adaptive dimensions of motivation are positively correlated with between-network constructs (academic satisfaction and adjustment); while impeding and maladaptive dimensions are negatively correlated with them. Taken together, these results support the conclusion that the concept of motivation and engagement defined by Martin is applicable to Chinese university students. To explore whether students can be grouped into different subtypes based on the MES-UC, Latent Class Analysis (LCA) was conducted. The results indicated that latent class analyses provided additional information that is not available from factor analysis. In particular, it identifies subgroups based only on engagement when factor analysis indicated that subgroups be identified on both engagement and motivation. These findings are discussed in the context of the need to conduct both content and psychometric analyses of questionnaire when they are used in cross-cultural settings. Keywords: academic engagement, MES-UC, latent class analysis, factor analysis.

P06-12

High stakes assessments and their impact as conceived by high-school teachers

Michaelides, Michalis (European University Cyprus/Open University Cyprus, Cyprus)*

Abstract

Studying conceptions about assessment is important because research has linked teachers' conceptions about teaching and learning with their practices. In Cyprus the educational system is not characterized by school accountability linked to assessment outcomes; most students participate in only one high-stakes, end-of-high-school national testing program for graduation and university entrance purposes. The objective of this paper is to investigate how lyceum teachers (10th-to-12th grade) conceive of the purpose of educational assessment, what role the national exams play in their assessment conceptions, and how they perceive the impact of exams on the high-school curriculum and education. Thirty purposefully sampled teachers of subjects tested, or non-tested in the exams were interviewed on the following thematic axes: conceptions about educational assessment, beliefs about the purpose of the national examination system, and experiences about the influence of the system on high-school education. A third of the transcribed interviews will be recoded independently to ensure reliable coding. It is hypothesized that lyceum teachers will emphasize assessment for imposing student accountability, while assessment as a means to improve teaching and learning will be less prominent. Findings will be contrasted with experiences and beliefs of students who have taken these high-stakes exams on issues like teaching to the test, the autonomy vis-à-vis dependence of secondary from tertiary education, and the dominance of tested subjects over other curriculum areas. The investigation is timely as the local authorities are currently debating reforms in the examination and assessment practices in the Cypriot public education.

P06-15

Assessment of the needs on evaluation of different schools in the Philippines

Parinas, Neil (De La Salle-College of Saint Benilde, Philippines)*
Mamauag, Maria Felicitas (Marife) M. (De La Salle-College of Saint Benilde, Philippines)
Magno, Carlo De La Salle University, Manila, Philippines)

Abstract

The study assessed the evaluation needs of 81 schools nationwide. Using a descriptive research design, a survey questionnaire was constructed and validated by measurement experts. Descriptive statistics were obtained to describe the evaluation activities, practices and areas on evaluation needing technical assistance. Findings show that all schools conduct evaluations mostly on teacher performance (n=29.36%). There was an expressed need for technical assistance practically from planning to utilization of evaluation results, especially in areas of instrumentation and data analysis which were rated highly. The t-test results further indicate no significant difference in the needs between public and private schools. Implications and recommendations related to improve effective training programs that were drawn are discussed in the study.

P06-17

Validation of the French version of the multidimensional scales of perceived academic efficacy: A structural and a longitudinal study.

Vrignaud, Pierre (Université Paris Ouest Nanterre La Défense, France)*
Blanchard, Serge (INETOP)
Rocher, Thierry (Université Paris Ouest Nanterre La Défense, France)

Abstract

To study the perceived academic efficacy of 11 to 16 years pupils for a longitudinal survey of French pupils, we realized a French adaptation of the the Multidimensional scales of perceived academic efficacy (Bandura, 1990). The survey is a follow up during four years of a representative sample of 30,000 pupils. The first step consists in a validation study to get information about the psychometrical validity of the French adaptation and to compare the structural validity of this version with data gathered in different European countries (Pastorelli, Caprara, Barbaranelli, Rola, Rozsa, & Bandura, 2001). The second step is a longitudinal study (the same pupils observed two years later) aiming at identifying the change in the perceived academic self efficacy particularly in its structure: are the different academic domain more distinct as the pupils are getting older and advancing in their curriculum. The numerous variables gathered in this survey allow to test causal hypothesis on the reciprocal effects of the perceived academic self efficacy and the pupil's results. These results will be discussed to highlight the advantages and drawback of the French version of the Multidimensional scales of perceived academic efficacy.

P06-18

The development of teachers' situational judgment test based on the critical teaching behavior events

Xu, Jian Ping (School of Psychology, Beijing Normal University, China)*
Tan, Xiao Yue (School of Psychology, Beijing Normal University, China)
Wu, Lin (School of Psychology, Beijing Normal University, China)
Yang, Min (School of Psychology, Beijing Normal University, China)

Abstract

It is necessary to develop a set of teachers' situational judgment test in China due to the teachers' competency tests currently used in China are mainly self-report questionnaires being vulnerable to social desirability. The autobiographies of the 15 excellent teachers instructing in primary & middle school were analyzed by quality analysis methods. Critical behavior events happened in each teacher's teaching experiences and their ordinary teaching life were collected and analyzed by critical incidents technique. Competencies acquired earlier in previous research were main reference for analyzing the critical incidents. Representative educational scenes were chosen and the corresponding critical incidents were edited into situational stimuli keeping similarities in length and form. Expert teachers and novice teachers were asked to generate responses for each situational stimulus. Another group of expert teachers were asked to rate the efficiency of each response to form response choices and record keys and form the teachers' situational judgment test for pencil test and interview. Test the teacher applicants and carry on the follow up study keeping observing the employed teachers' teaching performance to explore the validity of the test items.

P06-19

Research of sources and causes on various kinds of university students' psychological stress

Yang, Min (Shanghai International Studies University, China)*

Abstract

In order to explore the sources of psychological stress in various kinds of university students and analyze the main causes of university students' stress, this article approached self-developed Psychological Stress Scale of University Students and 224 undergraduates completed the questionnaires. The final result shows during 4 years of study, stress of freshmen is higher than the other grades; except learning stress, the other stresses are of great discrepancy in students from different grades; stress of students with different gender and from different sources differ a lot.

P06-20

A standardized test in Arabic measuring level of child performance

Zellal, Nacira (Laboratory of Language Sciences – Cognitive Neurosciences – Communication (SLANCOM), Algiers University, Algeria)*

Abstract

This is the presentation of the first Algerian psycho-clinical full battery of 33 tests for the evaluation of child performance in communication process. It represents the results of an Algerian - French Project of Cooperation (1991-1996). The original version of this battery is called the "MT 86" , edited by OrthoÉdition, Icebergues , France (1992), has been conceived for francophone aphasic adults neuropsychological approach. However, a long empiric experience with the same battery adapted to Algerian culture, shows its efficiency in child acquisition deficits as dysphasia, delay of language and word, dyslexia and dysorthographia, grammatical and arithmetic troubles. First, I shall present the procedure of: 1) its adaptation to the Algerian psychosociolinguistic context knowing that adapting is not translation. Concerning linguistic systems of adaptation operation, oral and written Arabic, then berbere languages are experimented; 2) its standardization upon 460 normal subjects. Then, through the projection of selected fragments of a CD Rom presenting the battery itself, I shall show the practical usefulness of this clinical instrument. Last, I shall show that when using French psycho-clinical tests since the independence of Algeria (1962), the clinician risks to give wrong diagnosis of child language and communication disorders. Case studies examples will be synthesized during the oral presentation.

P06-21

Appropriate application of criterions in selecting multilevel latent class model

Zhang, Jie-ting (Psychological Application Research Center, South China Normal University, China)
Zhang, Min-qiang (Psychological Application Research Center, South China Normal University, China)*

Abstract

Determining the number of class in mixture model is of great significance for latent class analysis, which directly decides the parameters of model and affects the accuracy of classification. When selecting multilevel model through the usually used criterions, such as BIC, AIC or AIC3, more factors should be considered about. Vermunt et al. have discovered in their research that these criterions have different advantages in different contexts of sample size. The present study, on basis of some conclusions made in the previous studies, is to investigate how sample size in individual level and group level influence the accuracy of model selection using these criteria when the latent class number in group level is relatively large, ranging from 4 to 6, the result of which is also compared with the former one. The stimulation result is expected to conform to the result in Vermunt's study: 1) BIC is good at judging the number of mixture contribution when sample sizes in group level and individual level are both large enough; 2) AIC has better judgment than others when sample sizes in the two levels are small; 3)unlike BIC or AIC, AIC3 dose not suffer from underestimation or overestimation of class number when sample size is small or large; consequently, AIC3 is quite a good criterion in general. Combining with the results of previous studies, an overall summary is made for proper use of these criterions. Keywords: multilevel latent class model, criterion in model selection, sample size, number of class

P07 08:30-10:00 Room105

P07-1

Reliability, validity and factorial analysis of the Anger Rumination Scale

Besharat, Mohammad Ali (University of Tehran, Iran)*

Abstract

The main purpose of this study was to examine the psychometric properties of Farsi version of the Anger Rumination Scale (ARS) including reliability, validity, and exploratory factor analysis. Four hundred and fourty nine students from the University of Tehran (234 females, 215 males) were included in this study voluntarily. All participants were asked to complete the ARS, the Tehran Multidimentional Anger Scale (TMAS), and the Mental Health Inventory (MHI-28). The results of explarotary factor analysis supported four factors for the ARS as well as a single general factor of ruminatiom. Content validity of the ARS was calculated according to Kendall's coefficients of concordance for ARS subscales including Angry Afterthoughts, Thoughts of Revenge, Angry Memories, and Understanding of Causes as well as ARS total score. All Kendall's coefficients of concordance were statistically significant. The convergent and discriminant validity of the ARS were supported by an expected pattern of correlations between the ARS and the measures of trait-anger, state-angerand, anger-in, anger-out, anger-control in, anger-control out, psychological well-being, and psychological distress. Test-retest reliability and internal consistency of the ARS were also examined at satisfactory levels. It was concluded that the ARS is a reliable and valid scale to measure anger rumination.

P07-2

Immigrant status – same definition, different meaning. Comparative analysis of immigrant status in PISA 2006 in Croatia and UK

Curkovic, Natalija (National Centre for External Evaluation of Education, Croatia)*
Sabic, Josip (National Centre for External Evaluation of Education, Croatia)
Buljan Culej, Jasminka (National Centre for External Evaluation of Education, Croatia)
Elezovic, Ines (National Centre for External Evaluation of Education, Croatia)

Abstract

International studies of school achievement mostly use the immigrant status as an important variable in a prediction of student's success. In PISA, immigration background is based only on three variables: students' and their mother's and father's country of birth. The aim of this paper is to compare actual meaning and functioning of the immigration background variable in two different national contexts using the PISA 2006 Croatian and UK data. Structure of immigration background variable was analyzed and compared between countries and with the official immigration statistics in both countries. Unexpectedly, results showed that in Croatian sample were more immigrants than in the UK sample. An analysis of nationality structure indicated that in Croatian sample immigrants are dominantly Croatian people from Bosnia and Herzegovina whose mother tongue is Croatian. Most of them have Croatian national identity and Croatian citizenship so they do not see themselves as immigrants, while in UK sample immigrants are dominantly from countries with different culture and language. In spite of the same definition of immigrant status for all participating countries in PISA, meaning of the obtained results is not the same. Definition of immigrant status in PISA seems more appropriate for western countries in which immigrants, who came from different cultures and speak different language, are dominant group. Mother tongue, time spent in host country, immigration policy and cultural differences between country of origin and host country must be taken into account when using immigration status as a predictor variable.

P07-3

Psychometric evaluation of the Italian translation of the "Assessment for Adults with Developmental Disabilities" questionnaire

De Vreese, Luc Pieter (Psychogeriatric Service, Health District, Modena, Italy)
De Bastiani, Elisa (ANFFAS Trentino Onlus, Italy)*
Mantesso, Ulrico (ANFFAS Trentino Onlus, Italy)
Gomiero, Tiziano (ANFFAS Trentino Onlus, Italy)

Abstract

Dementia-related maladaptive behaviours and neuropsychiatric symptoms in ageing people with intellectual disabilities (ID) are frequent. The AADS questionnaire rates operationally defined observable dementia-associated Behavioural Excesses and Deficits on the basis of their frequency of occurrence, care management and quality of life effects on the individual concerned, allowing for person-centred intervention procedures and for the verification of their efficacy also from the perspective of the patients themselves. The internal consistency and intra- and inter-rater reliabilities of the AADS-I was tested on a sample of 63 adult subjects with Down's syndrome (DS) and other forms of ID, attending day and residential services of the Province of Trento. All of the ID subjects were without severe sensory and language deficits, recent stressful life events, and clinically relevant psychiatric and organic comorbidity. All Cronbach's alpha coefficients were above the 0.70 criterion, inter-rater reliability was satisfactory for the six sub-scales with coefficients raging from 0.67 to 0.79., whereas interclass correlations for the intra-rater reliabilities ranged from 0.71 to 0.81. The frequency of occurrence of both the behavioural excesses and deficits was not significantly associated with age and gender, whereas only the deficit frequency sub-scale correlated with the level of ID severity. The dementia subgroup manifested a more frequent occurrence of behavioural deficits as compared to subjects without clinically manifested dementia, after controlling for age, gender and type and severity of pre-morbid ID.

P07-4

The Italian version of Intellectual Disabilities-Caregiver Difficulty Scale (CDS-ID). An instrument for measuring the perceived difficulties in the care of people with intellectual disabilities

Gomiero, Tiziano (ANFFAS Trentino Onlus, Italy)
Dalmonego, Carlo (ANFFAS Trentino Onlus, Italy)
De Bastiani, Elisa (ANFFAS Trentino Onlus, Italy)*
Weger, Elisabeth (ANFFAS Trentino Onlus, Italy)

Abstract

The aim of this study is the Italian validation of the Caregiver Difficulty Scale (CDS-DI), an instrument for assessing perceived difficulty by the caregiver for adult/elderly person with intellectual disabilities (ID). There is a shortage of instruments for measuring perceived burden by formal caregivers, while it may be predicted in more restrictive settings for persons with ID and primary or secondary dementia. The research involved 58 staff caregivers of persons with ID and findings showed a good internal consistency with Cronbach's ; ranging from 0.87 to 0.96, with item-total correlations of all sub-scales higher than 0.49 and good intra-rater reliability (0.77). These results show that this new scale provides a good measurement of the perceived difficulty and offers an opportunity to highlight in a systematic way the experiences of staff who cares for persons with ID with or without dementia.

P07-5

Development of the diagnostic dyslexia test of pronunciation and font style based on CTT and RSM

Fan, Xiao ling (Hunan Normal University, China)*
Liang, Juan (Hunan Normal University, China)

Abstract

The difficulties which ranged from 0.50 to 0.89 from test A items is 93.7%and test B items is 84.4%,81% discriminational index is above 0.30. The parallel form reliability is 0.77, the retest reliability of A, B are 0.88 and 0.96, cronbach' coefficient is 0.91 and 0.92. The correlation between the total score of A, B and 3-4 grade reading understanding examination are 0.64 and 0.74, correlation between the total score and chinese are 0.80 and 0.76. Taking 21 as cutoff score, the accuracy probability and detection rate of test A are 0.87;0.81, and those of test B are 0.83;0.60.Finally, there are six subtests enter the rule space analysis, respectively the Phonology, the tonality, the syllable, the pronunciation and meaning of the character which is the same shape but different sound and the mirror image character. The dyslexic child population is the most in the typical grasp pattern 1, 2, 3, 4, 5, 7, 10, A and B occupy above 80%. The dyslexic child has the pronunciation processing barrier. In the glyph, besides the meaning which character is the same shape but different sound, the child grasp below 50% on the mirror image character and the pronunciation of character is the same shape but different sound. Conclusion: Test construction for diagnosing pronunciation and glyph dyslexia of elementary student meets the requirement of metrology, its pattern of typical attribute grasps and the probability of attribute supply the effective information to diagnose for the reading dyslexic child.

P07-7

Dementia questionnaire for persons with intellectual disabilities (DMR): A reliability study of an Italian version

De Vreese, Luc Pieter (Psychogeriatric Service, Health District, Modena, Italy)
Mantesso, Ulrico (ANFFAS Trentino Onlus, Italy)
Gomiero, Tiziano (ANFFAS Trentino Onlus, Italy)*

Abstract

This study aimed to verify the reliability of the Italian version of the DMR, an informant-based questionnaire aimed at screening for dementia in persons with intellectual disability (ID). Methods: the DMR has been administered to the informants of 60 persons who attend the services of four of the principal organisations specialised in ID of the Province of Trento. DMR's internal consistency and intra-rater and inter-rater reliability were calculated. Results: the persons with ID who have been evaluated with the DMR have a mean age of 51,2±6,3 SD years, 29 persons are females, 51,6% has a Down syndrome, and mean QI is 29,1 ± 13,2 SD. Homogeneity of the DMR is found to be excellent for the Sum of Cognitive Scores (SCS) and good for the Sum of Social Scores (SOS) with Cronbach's a of 0.92 and 0.76, respectively. Interclass correlation coefficients of 0.93 and 0.91 for the SCS, and of 0.84 and 0.82 for the SOS, demonstrate excellent test-retest and inter-rater reliabilities. QI values, but not age and gender, correlate with the DMR. The DMR results to be applicable to persons with profound ID. Conclusions: these results confirm the reliability of the Italian versions of the DMR.

P07-8

Sources of hope in China and the Philippines: A cross-cultural validation of the Locus of Hope Scale

King, Ronnel (The University of Hong Kong, Hong Kong SAR, China)*
Bernardo, Allan (De La Salle University, Philippines)
Du, Hongfei (The University of Hong Kong, Hong Kong SAR, China)

Abstract

Current research on the hope construct has used an individualistic definition emphasizing personal and direct agency. A more collectivistic definition of hope that taps into indirect and/or proxy agency and how significant others play a role in hopeful thinking is needed, thus a new measure of hope called the Locus of Hope Scale (LHS) was developed that tapped into how friends, parents, and spiritual being(s) play a role in the hope construct. The aim of this study was to establish the construct comparability and cross cultural utility of the English and Chinese versions of the Locus of Hope Scale (LHS). Based on data from 200 Chinese and 200 Filipino university students, multi-group confirmatory factor analysis revealed the validity of the proposed factor structure across the two cultures. This research supports the use of the LHS in cross cultural research and provides data about a broader range of hope than has been investigated previously.

P07-9

Measurement of hope in Chinese adolescents: The Chinese Children Hope Scale

Leung, Beeto (University of Hong Kong, Hong Kong SAR, China)*
Ho, Samuel (University of Hong Kong, Hong Kong SAR, China)

Abstract

Hope is a cognitive, emotional and motivational construct that has been found to associate with many positive outcomes in Western youth. However, little evidence has been reported about the reliability, validity and factor structure of the Children Hope Scale, used to measure hope, in Chinese samples. This study examined the psychometric properties and factor structure of the Chinese version of Children Hope Scale (C-CHS). The C-CHS, along with other measures assessing their general self-esteem, happiness, satisfaction with life and depression, was administered to 902 Chinese secondary school students. Results showed that the C-CHS was internally consistent and temporally stable across 9 months. Regarding the concurrent validity of the C-CHS, the total scores together with the pathway and agency subscores were all significantly and moderately correlated with higher self-esteem, higher level of happiness and satisfaction with life and lower level of depression. Consistent with the original study, confirmatory factor analysis revealed empirical supports to the 2-correlated-factor conceptual model over the 1-factor model of CHS. The study has extended supports for the reliability, validity and the underlying factor structure of the CHS in Chinese sample. The applicability of C-CHS in Chinese samples and directions of future studies were discussed.

P07-10

The development of the Native Multicultural Literacy Scale

Lin, You Zhen (National University of Tainan, Taiwan)*
Twu, Bor-Yaun (National University of Tainan, Taiwan)

Abstract

Recently, school teachers in Taiwan have had more opportunity to work with multi-ethnic students and parents. Multicultural literacy is necessary for them to provide culturally responsive teaching and to make good cultural communication. However, some empirical studies show that most teachers are not yet prepared to practice multicultural education. The Multicultural Literacy Scale could help teachers evaluate their multicultural competence and the multicultural training courses they attend, as well as help students evaluate their school environment. In a word, the Multicultural Literacy Scale with multiple functions could help teachers promote their multicultural literacy. Based on literature review, this article firstly argues the importance of developing instrumentation for evaluating teachers' multicultural literacy; and then a native multicultural literacy scale will be developed and the Rasch model will be used to analyze data obtained from this scale. Reliability and validity evidence will also been discussed.

P07-12

Validity on S-EMBU-A: Comparison of results between classical and SEM approaches in CFA for repeated measures

Penelo, Eva (Departament de Psicobiologia i Metodologia, Universitat Autònoma de Barcelona, Spain)*
Viladrich, Carme (Departament de Psicobiologia i Metodologia, Universitat Autònoma de Barcelona, Spain)
Domènech, Josep M. (Departament de Psicobiologia i Metodologia, Universitat Autònoma de Barcelona, Spain)

Abstract

This study aimed to compare the results of validity obtained with two techniques in the framework of confirmatory factor analysis (CFA): classical approach vs. structural equation modeling (SEM). 284 Spanish psychiatric outpatient adolescents answered S-EMBU-A (a 22-item and 3-factor questionnaire assessing perceived parental rearing practices) during a clinical assessment. Two steps were performed: 1) measurement invariance across father's and mother's ratings was established with CFA for repeated measures; 2) covariates were added to the fully invariant model to test the influence of sex and age on latent means, and correlations between latent means and measured variables of family functioning were established. Classical approach with ANOVA (sex*age) and bivariate correlations also examined the relationship between scores and the external variables. Most of the CFA parameters were equivalent across father's and mother's ratings. Latent means for Emotional Warmth and Overprotection were higher for mother's ratings, the latter disappearing once the effects of covariates and structure coefficients were added. Influence of age found with ANOVA was non-significant in SEM. The mean difference between bivariate correlations and standardized parameter estimates for the SEM model was ;0.04;. The pattern of relations remained essentially unchanged when comparing SEM and the classical approach to examine validity. However, SEM approach requires more precision on model specification.

P07-13

The PRT – A Picture Based Test to study the development of mental rotation

Rohe, Anna (Universität Koblenz-Landau, Campus Koblenz)*
Quaiser-Pohl, Claudia (Universität Koblenz-Landau, Campus Koblenz)

Abstract

The Picture Rotation Test (PRT) was designed to explore the mental-rotation ability of pre-school children at the age of four to six. The test consists of 16 images of humans and animals. For each image the child is asked to compare it to three other images and decide whether they are the same image or a mirror image. In contrast to mental-rotation tests for adults the images are only rotated in the plane (not in depth). As of the pictorial quality of the test and its few verbal instructions, we expect good applicability for studying the development of mental rotation with children from different cultures. The developmental stage of mental rotation can be identified via analyzing the solution strategies in the PRT with the statistical method of latent-class-analysis. This approach has been applied with a sample of n=565 German pre-school children, results will be reported and discussed.

P07-14

Neckers cube drawing in screening for cognitive dysfunction in clients with various ethnical and educational background

Sundseth, Øyvind (NPF, Norwegian Psychological Association, Norway)*
Marberger, Tove Kanestrøm (NPF, Norwegian Psychological Association, Norway)

Abstract

We studied a group of 75 long-term social assistance recipients, where 41% had a minority background (refugees/immigrants). Traditional screening procedures are often not very well suited for use in such client populations. The copying of Neckers cube is often regarded as a procedure sensitive for brain dysfunction, relative robust against influence from factors as intelligence, age and psychiatric illness. It is also easy administered and is relatively effortless for the client to complete. It is documented that 95 % of children adapt the skill to copy this figure before the age of 12, and that the skill is kept unchanged in adults well into their sixties (Strub et al 2000). We expected that the copying of Neckers cube would be less sensitive to milder cognitive problems in our group, but hopefully more specific in indentifying marked/serious problems. The goal is to identify individuals with a disabling cognitive dysfunction in a group where presumably most of the participants have a cognitive problem to some degree, and to validate the drawing of Neckers cube as screening procedure for spesific cognitive dysfunction using the WAIS-III as the gold standard. Preliminary Results: The incorrect drawing of the Neckers cube identified 17 clients with possible serious cognitive dysfunction, 16 showed persistent copying problems over 3 trial, while WAIS-III screening confirmed that these persons had one or more marked cognitive dysfunctions. Preliminary Conclusion: The Neckers cube identified 17 persons with marked copying/drawing problems. An extensive cognitive evaluation confirmed a major cognitive dysfunction in 16 of these clients. The drawing procedure could have the potential to be a low cost and low effort screening procedure in identifying cognitive problems. Thus we conclude that the Necker cube drawing had a high specificity with regard to identifying serious cognitive dysfunction in our study. Still further studies are needed to confirm its utility.

P07-15

Development and application of mandarin Stroke Impact Scale in Taiwan

Huang, Yuching (National University of Tianin, Taiwan)
Wu, Chiaoying (National University of Tianin, Taiwan)*
Chang, Kuchou (Department of Neurology, Chang Gung Memorial Hospital, Kaohsiung, Taiwan)

Abstract

There is no Taiwanese scale available to assess quality of life after stroke. Translation of scales developed by English speaking areas is used instead. However, translation of the scales might be limited due to cultural or demographics differences. This study was designed to test the reliability and validity of a mandarin version of American Stroke Impact Scale version 3. By exploratory factor analysis and component analysis, reducing the items of the mandarin scale is attempted. This mandarin version had fair reliability of internal consistency 0.94. exploratory factor analysis indicated 4 components factors including 54 items with total explanation 68.9% but not related to the original scale for the quality of life. This mandarin version had adequate power of explanation. It might help to alleviate the burden of examiner, facilitate the assessment of the quality of life after stroke. Further large scale tests validating this mandarin version are anticipated.

P07-16

The structure of self-esteem: Single dimension or two dimensions?

Xu, Jian Ping (School of Psychology, Beijing Normal University, China)*
Wu, Lin (School of Psychology, Beijing Normal University, China)

Abstract

Rosenberg produced his Self-Esteem Scale based on the single dimension assumption of self-esteem. In recent years, many researchers who used this scale had found that the structure of self-esteem is not a single dimension. They further proposed a self-esteem structure of two dimensions. This study try to find out if the structure of self-esteem is single dimension or two dimensions, and if there is a cross-cultural stability of the structure of self-esteem in a variety of ages, genders, occupations and educational levels in Chinese cultural background. Subjects randomly selected from 9 to 60 years old groups, who were grown up in Chinese culture background. After using the Rosenberg's Self-Esteem Scale, this study discussed the internal structure of self-esteem by the approaches of exploratory factor analysis and confirmatory factor analysis. The results showed that: (1) in this scale, there were four negative items loading on one factor, meanwhile six positive items loading on the other factor, which were consistent with the results of previous studies; (2) with the age increased, there were no significant changes showed up in their SES's scores; (3) there were no significant differences in the SES's scores among those people in different occupations, gender, educational backgrounds; (4) there was no significant difference between the average of Chinese adults' scores and the average of foreign norm's scores. Study concluded that: (1) there was a two-dimensional structure measured by Rosenberg's Self-Esteem Scale; (2) the existence of cross-cultural stability has been found in different ages, genders, occupations and educational levels.

P07-17

Measurement invariance and construct compatibility

Yang, Xin Sophie (Loughborough University, United Kingdom)*
Jowett, Sophia (Loughborough University, United Kingdom)

Abstract

Coach-athlete relationship has attracted the interest of researchers from all over the world. One important feature of international research is to ensure that the measurements used are reliable and valid both within and between diverse cultural contexts. Coach-Athlete Relationship Questionnaires (CART-Qs) have been developed to assess the quality of the relationship as this is defined by the constructs of Closeness, Commitment, and Complementarity (3Cs) from both coaches' and athletes' points of view. The present study primarily focused on testing for the equivalence of the items contained within the athlete version of the CART-Q, as well as the underlying theoretical multidimensional structure of its three subscales across seven countries, namely, United States of America (N =177) Belgium (N =200), Britain (N =382), China (N =200), Greece (N =115), Spain (N =120), and Sweden (N =169). Multi-group mean and covariance structures analysis (MACS) were conducted to test for the measurement and construct equivalence of the CART-Q within and across cultures. Results revealed evidence of similarly specified factor structures of the 3Cs for all cultural groups as well as partially measurement and structural invariance. Overall, findings provided support for the psychometric properties of the CART-Q in diverse cultures. The significant differences in latent means of the CART-Q across groups suggest that there may be cultural variation in terms of the intensity athletes perceive the quality of the coach-athlete relationship. The possibility of subtle cultural variations in the multidimensional conception of coach-athlete relationships are discussed alongside implications for theory, research, and practice.

P07-18

Psychometric properties of the Gratitude Questionnaire -6 (GQ-6) in Iranian adolescents

Ghamarani, Amir (Dept of psychology, University of Isfahan, Isfahan, Iran)*
Kajbaf, Mohammad, B (Dept of psychology, University of Isfahan, Isfahan, Iran)
Oreizi, Hamid. R (Dept of psychology, University of Isfahan, Isfahan, Iran)
Amiri, Shole (Dept of psychology, University of Isfahan, Isfahan, Iran)

Abstract

The aim of this study was to investigate psychometric properties of the Gratitude Questionnaire -6 (GQ-6,Immond, McCullough & Tsang, 2002)) in a sample of Iranian Adolescents aged 14 to 17 years (n= 200). The factorial analysis showed that the Iranian scale has the same factorial structure as in the original version of the scale with one factor. The Iranian Gratitude Questionnaire -6 (I- GQ-6)exhibited satisfactory internal and test – retest coefficient. Moreover, the girls' superiority in gratitude was replicated. So, The I- GQ-6 is one of the first gratitude scale to be made available in Iran and possesses good psychometric qualities.

P07-19

The impact of teacher variables on the fourth-grade math achievement: A value-added study

Zhang, Wenjing (Beijing Normal University, China)*
Xin, Tao (Beijing Normal University, China)

Abstract

The subjects of the study were 1238 fourth-grade students and 42 math teachers from 42 primary schools in Fangshan District, Beijing. The fundamental value-added model was constructed by adding the math achievement of the last semester as the covariate and the counterpart of the second semester of the fourth-grade as the dependent variable in the two-level linear model. Then on the condition of controlling students' background information, we examined the relationship between three aspects of teacher variables (teacher characteristic variable, teaching methods and teacher training) and the gain of students' achievement and discussed the differences between level indicator and value-added indicator of teacher evaluation. The results indicated: teachers' gender, age, the year of teaching and major had no significant effect on the achievement gain, while teachers' professional title, final educational qualification and whether teachers had participated for the new curriculum training had significant effect on the achievement gain. Moreover, there were great differences between the value-added indicator from value-added assessment method and the level indicator (e.g., average). Compared to the level indicator, the value-added indicator could provide more up-to-date, precise and ample information.

P07-20

Research on the attitude of Normal University students to homosexuality

Lv, Shaobo (Institute of Psychology, Chinese Academy of Science, China)*

Abstract

To explore the factors which affect Normal University students' attitude toward homosexuality. Questionnaires were distributed to Normal University Students. A total of 621 participants completed the questionnaires. Of these, 15 were homosexuals, 621 were heterosexuals with 144 males and 477 females. On the adapted version of Attitude Toward Lesbians and Gay Men Scale, boys had more scores than girls, and subjects had more scores toward male homosexuality than female homosexuality, and Singletons had fewer scores than the others. Attitudes toward sex, other ways of sexual intercourse in addition to penis-vagina mode are positively correlated to attitude toward homosexuality and attitude on social morality is negatively correlated to attitude about homosexuality. The difference of attitudes toward homosexuality were not significant for the four gender role types of males while significant for those of females which showed that feminine girls had more scores than masculine girls on the adapted version of Attitude Toward Lesbians and Gay Men Scale. Conclusions: 1. Contrast to girls, boys hold more negative attitude to homosexuality. 2. Subjects have more negative attitude toward male homosexuals. 3. Singleton have less negative attitude than students who have brothers or sisters. 4. Attitude toward sex and attitude on other way of sexual intercourse in addition to penis-vagina mode are positively correlated to attitude toward homosexuality. 5. Feminine girls have more negative attitude on male homosexuality than masculine girls.

Index of Contributors

General Information

Transportation

Coming to CUHK

If you are staying in the Hyatt Regency Shatin Hotel, the conference venue at the Chinese University of Hong Kong is within walking distance.

If you are staying in other hotels in Shatin (Royal Park Hotel and Regal Riverside Hotel), you can go to the Shatin MTR Station, and take the north-bound train to the University Station (the second stop from Shatin). The trip takes about 15 minutes including waiting time, and the fare is HK$4. You may also choose to take a taxi. The taxi fare is about HK$60 and the trip takes about 8 minutes.

If you are staying in hotels in Kowloon, take the MTR East Rail Line to the University Station. The trip takes about 30 minutes. You may also choose to take a taxi to the conference venue.

All meetings will take place in the Hotel Teaching Building, which is right next to the Hyatt Regency Hotel on Chak Cheung Street. This Hotel Teaching Building stands outside of the main campus. If you are taking the MTR East Rail Line, take Exit B from the University Station (direction toward Hyatt Hotel). If you are taking a taxi, please ask the driver to go the University Station of the MTR East Rail Line (instead of to the Chinese University of Hong Kong).The Hotel Teaching Block is behind the taxi rank.

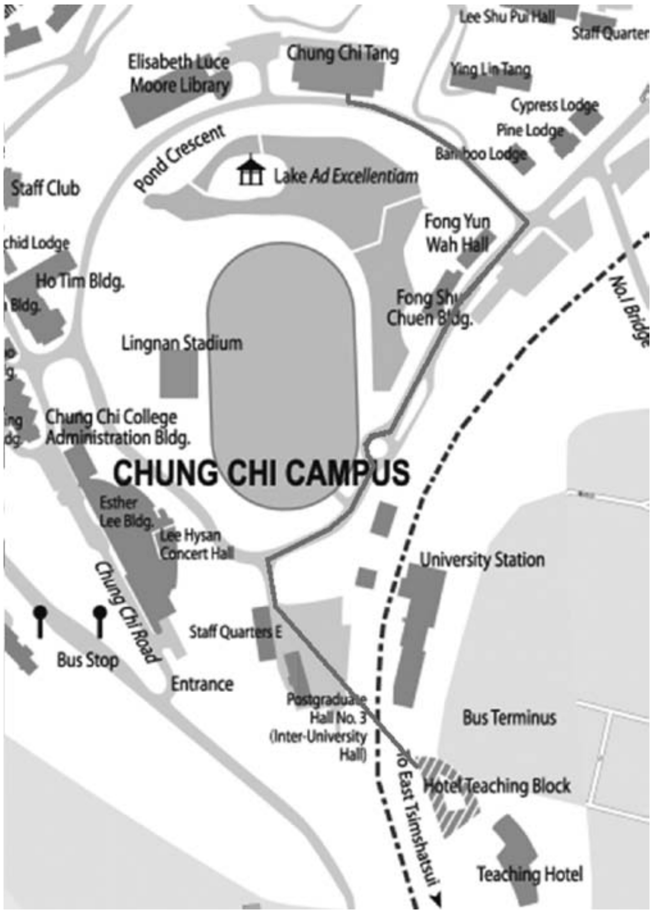
For enquiries about the conference, please call (852) 2609-8145

Tea break arrangement

Coffee, tea and cookies will be served at the morning and afternoon tea break on Level 2.

Lunch arrangement

Simple lunch boxes will be served on Level 3 of the Conference venue on Monday (19 July 10) and Tuesday (20 July 10) for participants. Lunch coupons will be put in your registration packet. For students, lunch will be provided in Chung Chi Tang, in 5-8 minutes' walking distance. Please refer to the following map for the route from the conference venue to Chung Chi Tang.



Location of water fountains

Distilled water will be served at the water fountains located on Level 1 and 2.

Internet Access

There will be free wireless Internet access at the conference venue, for which you will need to bring your own computer and obtain a password. To obtain a password, please go to the Computer Room on Level 2. The instructions on how to access the internet will be given to you when you get your password. Alternatively, there will be four desk top computers available for checking the files for presentations and accessing the Internet in the same room (the Computer Room on Level 2). Please note that each user is limited to 15 minutes' use of the computers at a time if other participants are queuing up. Priority will be given to participants checking their files for presentation. The hours of operation for the Computer Room will be: 8:30am to 5:30pm from 18 to 20 July 2010, and 8:30am to 1pm on 21 July. For the location of the Computer Room, please refer to the venue floor plan on page 5 in this book.

Contingency plan in case of inclement weather

The rules for cancellation or rescheduling are as follows:

a. In case Typhoon Signal No.8 or Black Rain Warning Signal is hoisted by 7am, the morning sessions will be cancelled. If the above signals are still hoisted at 12noon, the afternoon sessions will be cancelled.

b. If the workshops in the morning of 18 July (Sun) are cancelled, they will be rescheduled to the afternoon on the same day; cancelled workshops in the afternoon sessions would be rescheduled to the evening of 19 July (Mon).

c. If the programmes on 19 July (Mon) or 20 July (Tue) are affected, more parallel sessions will be arranged on the next two days.

d. If the programmes on 21 July (Wed) are affected, the cancelled sessions will not be rescheduled.

e. If the banquet on 20 July (Tue) is affected, it will be rescheduled to 21 July (Wed).

f. The announcement about the programmes that are rescheduled or cancelled in the event of inclement weather will be put on the website to notify conference participants.

Emergency number

In emergency situations, you can contact the local police, ambulance service, fire department and other emergency services by calling 999.

Hospitals

As an international city, Hong Kong has world-class hospitals providing outstanding care. Visitors using Accident and Emergency services in Hong Kong public hospitals are charged a set fee of HK$570 per visit, but will always be treated even if they cannot pay immediately.

The nearest hospital from the conference venue is:
Prince of Wales Hospital (with 24-hour Accident and Emergency Service)
Address: 30-32 Ngan Shing Street, Shatin, NT
Tel.: (852) 2632-22 11
Fax: (852) 2637-8244

General Health Advice for Participants

To ensure a healthy environment for all participants, please take note of the following precautionary measures from the University's Committee on Health Promotion and Protection if you have developed respiratory symptoms such as fever, malaise, chills, headache, joint pain, dizziness, rigors, cough, sore throat and runny nose:

- Put on a face mask when influenza symptoms or fever develop to reduce the chance of spreading the infection to other persons around you. A limited supply of surgical masks is available from the reception desk at the conference venue.
- Refrain from attending gatherings if fever or respiratory symptoms becomes severe.
- Observe good personal hygiene practices. Wash hands with soap and/or use alcohol handrub frequently. Cover nose and mouth with tissues when coughing or sneezing.
- Avoid visiting crowded or poorly-ventilated places.
- Avoid sharing of personal items among participants (e.g. towels, water bottles, etc.)
- Use serving spoons and chopsticks when sharing meals.

Notes

Notes

Diamond Sponsors:

Graduate Management Admission Council

Hogrefe-Verlag GMBH & Co. KG

Platinum Sponsors:

British Psychological Society

Law School Admission Council

Gold Sponsors:

ATA Ltd.

Pearson Language Tests

Educational Testing Service

College Board

Comms Multingual Ltd

Silver Sponsors:

Mobley Group Pacific Ltd.

American Psychological Association

Pearson Assessment

Acknowledgement

The Lion Dance at the Opening Ceremony is sponsored by the Hong Kong Tourism Board

Red and white wines at the Welcome Reception are sponsored by

DYNASTY
Since 1980

Dynasty Fine Wines Group Limited
王 朝 酒 業 集 團 有 限 公 司

Psychological Testing Centre
PTC
www.psychtesting.org.uk

The
British
Psychological
Society

The British Psychological Society's Psychological Testing Centre

The British Psychological Society is the leading organisation for setting standards in psychological testing in the UK.

The Society directs the work of its Psychological Testing Centre (PTC) through the Steering Committee on Test Standards whose role is to set, promote and maintain standards in testing.

The PTC website provides information and services relating to standards in tests and testing for test takers, test users, test developers and members of the public.

The PTC website offers you information on:

- Competence-based test user certification and registration in educational and occupational settings.
- Access to 140 test reviews in summary or full – reviewed against the *EFPA Review Model for the Description and Evaluation of Psychological Tests.*
- A list of tests which have met benchmark criteria for the award of a Test Registration Certificate.
- Guidelines and best practice statements on standards for the construction, use and availability of tests.

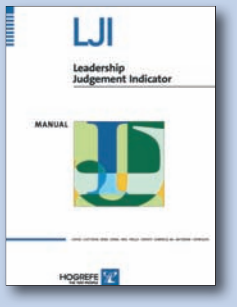
Contact the PTC:
Tel: +44 (0)116 252 9530
E-mail: enquiry@psychtesting.org.uk

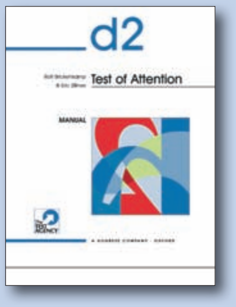
www.psychtesting.org.uk

International Assessment Tools

Leadership Judgement Indicator

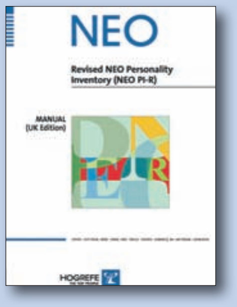
· English · German* · French · Danish* · Czech*





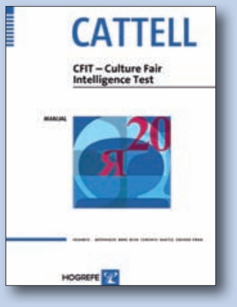
d2 Test of Attention

· English · German · French · Danish · Czech · Croatian · Spanish · Portuguese · Polish · Dutch · Hungarian* · Romanian



Revised NEO Personality Inventory

· English · German · French · Dutch · Danish · Czech · Swedish · Finnish*



Cattell – Culture Fair Intelligence Test

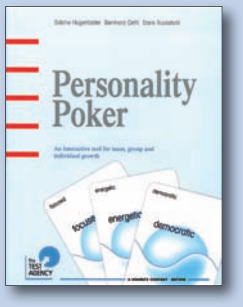
· English · German · Italian · Spanish · Croatian · Filipino* · Czech* · Danish* · Swedish* · French* · Dutch* · Polish*

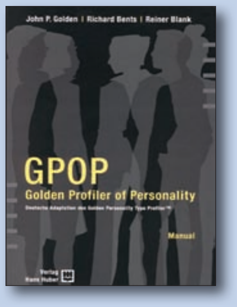
Intelligence Structure Test

· German · Czech · Italian · Lithuanian · Bulgarian · Slovak · English · French* · Dutch · Danish · Hungarian*

Personality Poker

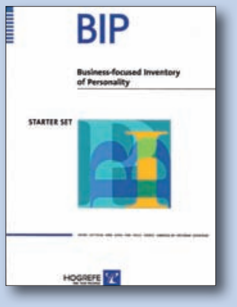
· English · German · Danish · Czech · Spanish · Finnish




John P. Golden | Richard Bents | Reiner Blank
GPOP
Golden Profiler of Personality
Manual

Golden Profiler of Personality

· German · Czech · English · French · Danish* · Slovak*



Business-Focused Inventory of Personality

· English · German · Czech · Spanish · Portuguese · French · Dutch · Danish · Slovak · Hungarian*· Croatian* · Bulgarian* · Finnish* · Brazilian Portuguese*

*(in press)

HTS International

The Hogrefe TestSystem (HTS)

The leading multilingual platform for computer- and web-based testing: flexible, secure, with tests in multiple languages, extensive tools for analysis and reports.

www.hogrefe-testsystem.com

SON-R Non-verbal Intelligence Test

for Children

· English · German · Dutch · French · Czech · Slovak · Danish/Swedish/Norwegian* · Romanian*

For a complete catalogue visit our Websites:

www.hogrefe.de
www.hogrefe.com
www.hogrefe.ch
www.hogrefe.fr
www.hogrefe.co.uk
www.hogrefe.nl
www.hogrefe.cz
www.dpf.dk
www.hogrefe.at
www.hogrefe.se

HOGREFE

Compliments of
Diamond Sponsor

GMAC®

GRADUATE MANAGEMENT
ADMISSION COUNCIL

International Test Commission (ITC)
Website: http://www.intestcom.or

Department of Psychology
The Chinese University of Hong Kong
Website: http://www.psy.cuhk.edu.hk/