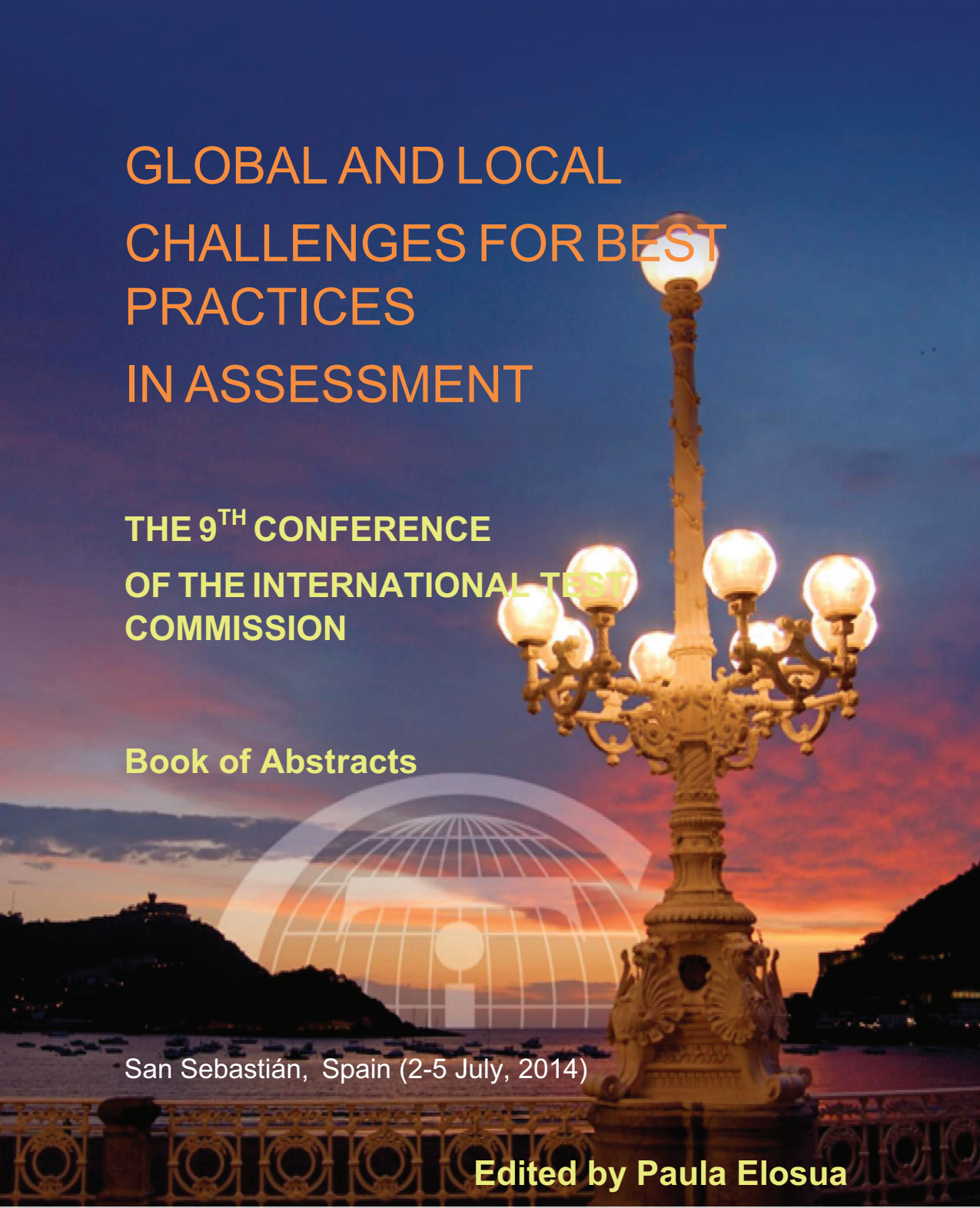


GLOBAL AND LOCAL CHALLENGES FOR BEST PRACTICES IN ASSESSMENT

THE 9TH CONFERENCE OF THE INTERNATIONAL TEST COMMISSION

Book of Abstracts

San Sebastián, Spain (2-5 July, 2014)

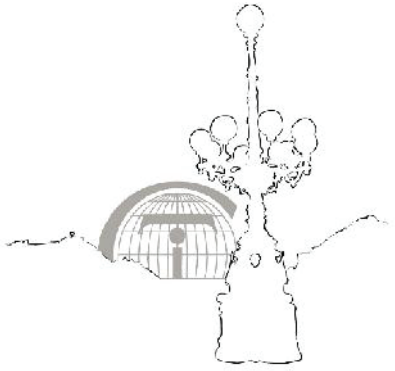
Edited by Paula Elosua

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

www.ehu.es/argitalpenak



The 9th Conference of The International Test Commission

Global and Local Challenges for Best Practices in Assessment

San Sebastian, Spain, 2 to 5 July, 2014

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

ARGITALPEN ZERBITZUA
SERVICIO EDITORIAL

Servicio editorial de la Universidad del País Vasco (UPV/EHU)
Euskal Herriko Unibertsitateko Argitalpen Zerbitzua

Press Service
ISBN: 978-84-9860-997-4
Legal deposit: BI-948-2014

Index

This work was partially financed by the Spanish Ministry of Economy and Competitiveness (PSI2011-30256) and by the University of the Basque Country (GIU12-32).

Presentation

Presentation

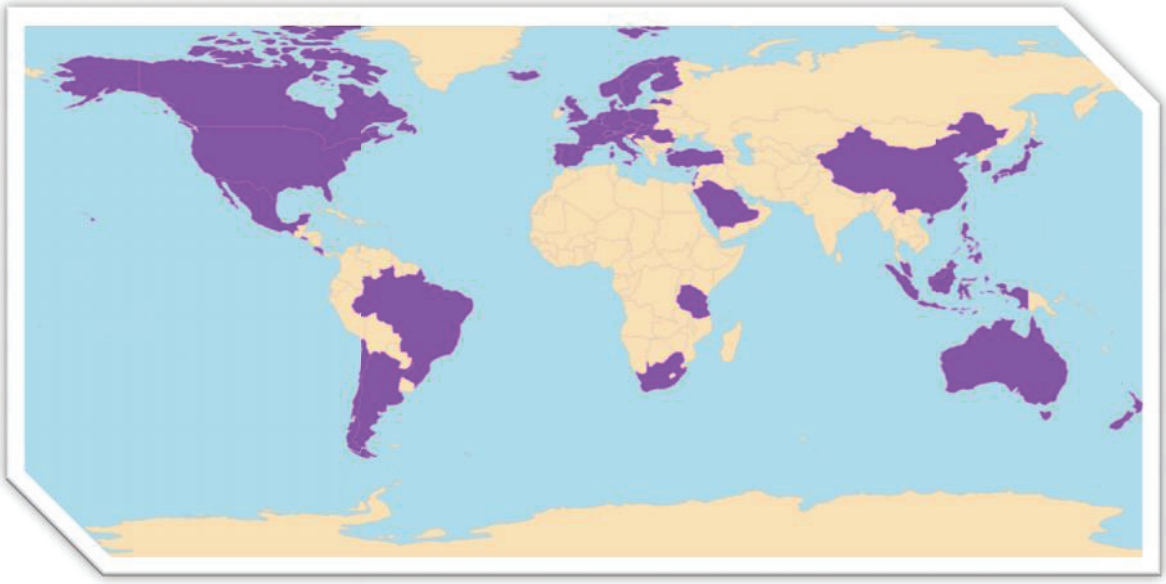
The 9th International Test Commission Conference (ITC) took place at the Miramar Palace in San Sebastian, Spain, between the 2nd and 5th of July, 2014. The Conference was titled, "Global and Local Challenges for Best Practices in Assessment."

The International Test Commission, ITC, is an association of national psychological associations, test commissions, publishers, and other organizations, as well as individuals who are committed to the promotion of effective testing and assessment policies and to the proper development, evaluation, and uses of educational and psychological instruments. The ITC facilitates the exchange of information among members and stimulates their cooperation on problems related to the construction, distribution, and uses of psychological and educational tests and other psychodiagnostic tools. The International Test Commission (www.intestcom.org) organizes biannual conferences. The previous Conference was held in Amsterdam, in 2012. The Conference in 2016 will be held in Vancouver, Canada.

This volume contains the Abstracts of the contributions presented at the 9th International Test Commission Conference. The four themes of the Conference were closely linked to the goals of the ITC:

- Challenges and Opportunities in International Assessment
- Application of New Technologies and New Psychometric Models in Testing
- Standards and Guidelines for Best Testing Practices
- Testing in Multilingual and Multicultural Contexts

In total, 501 presentations were made in 82 sessions. The Conference was attended by more than 500 professors and graduate students, educators, policy-makers, testing company representatives, and researchers from more



than 50 countries.

The presentations have been organized into nine sections:

- State of the Art Lectures
- Keynote Addresses
- Invited Symposia
- International Test Commission Special Sessions
- Symposia
- Round Tables and Pannel Sessions
- Oral Sessions
- Posters
- Workshops

This volume contains the abstracts as well as information about the authors and their institutions. An author index too is available to help readers locate contributions.

It has been a great honor for me to have the opportunity to organize this Conference and contribute to the development and dissemination of the work being carried out by the International Test Commission. I would like to offer my personal thanks to the invited speakers, to the organizers of the Symposia and the Workshops, to the members of the Scientific Committee, to the reviewers of the proposals, to the session Chairs, to the authors of all contributions to the Conference, and to the sponsors, for making the 9th International Test Commission Conference possible. Finally, we are greatly indebted to all of the attendees who came from more than 50 countries to participate.

Paula Elosua

Chair of the Organizing Committee

ITC President Welcome

It is my great pleasure to welcome you on behalf of the International Test Commission (ITC) and to invite you to attend the next International Test Commission conference to be hosted in San Sebastian, Spain, in 2014.

ITC is an association of national psychological associations, test commissions, publishers and other organizations, as well as individuals. ITC is committed to promoting effective testing and assessment policies and to the proper development, evaluation and uses of educational and psychological instruments. ITC facilitates the exchange of information among members and stimulates their cooperation on problems related to the construction, distribution, and use of psychological and educational tests and other psychodiagnostic tools.

Under the theme "Global and Local Challenges for Best Practices in Assessment", ITC 2014 will showcase new frontiers of assessment as a means for improving and developing psychological and educational tests and test uses across cultures, languages and national boundaries. This conference goes to the heart of all matters relating to testing and measurement, and brings together the best researchers and practitioners from around the world, allowing you to hear and meet those at the forefront of our discipline and practice. Following the success of a serie of 8 previous international conferences, which have been held in vibrant and fun locations in recent years, ITC 2014 offers an animated and personal atmosphere in which delegates hear the cutting edge developments in the field, and discuss how they resolve concerns common to our interests.

The location in San Sebastian combines culture, natural beauty, and enticing cuisine. There will be delegates not only from Europe, US, and other developed countries, but one feature is that delegates come from around the world -- Africa, South America, and Asia. In short, a wonderful place to meet up with one another.

I very much look forward to see you in San Sebastian in 2014.

Fanny M. Cheung

President, ITC

ITC2014 Organizer Welcome

Dear colleagues,

On behalf of the Organizing Committee of the 9th Conference of the International Test Commission, it is my great pleasure and privilege to invite you to join us in San Sebastian. After the very successful Conference in Amsterdam in 2012, the Conference comes for the first time to Spain. ITC2014 offer us the opportunity to combine the beauty and hospitality of San Sebastian with some of the outstanding work being done on the important topic of testing and testing practices, and to make easy contacts and relations among researchers from around the world.

San Sebastian ("Donostia" in Basque) is part of the Basque Country, with its amazing unspoilt landscape of mountains, waterfalls and beaches. San Sebastian is located around the beautiful La Concha Bay, and boasts one of the best in-city beaches in Europe. Our city is internationally renowned for being a culinary haven and its top-quality foods and mouth-watering "pintxos" or morsels are made available to visitors.

ITC2014 will be held at the Miramar Palace. This palace was constructed as quarters for Queen María Cristina and her entire court after her decision to take up summer residence in San Sebastian in 1885. The building offers impressive views over the two beaches adjacent to the centre of town and Santa Clara Island, directly opposite in the Bay.

We work closely with the Scientific Committee to provide an unforgettable scientific experience as well as to offer a social program filled with of opportunities to share the beauty of the Basque Country with colleagues and friends around the world in a very familiar and warm atmosphere.

I look forward to meeting together in the summer of 2014 for a great Conference in San Sebastian!

Paula Elosua

Chair of the Organizing Committee

Committees

Organizing Committee

- Paula Elosua (Chair, University of the Basque Country, Spain)
- David Bartram (SHL Group, United Kingdom)
- Fanny M. Cheung (The Chinese University of Hong Kong)
- José Muñiz (University of Oviedo, Spain)

Local Organizing Committee

- Mikel Garcia (University of the Basque Country, Spain)
- Josu Mujika (University of the Basque Country, Spain)
- Alicia López-Jaúregui (University of the Basque Country, Spain)

Scientific Committee

- John Hattie (Chair, University of Melbourne, Australia)
- David Bartram (SHL Group, United Kingdom)
- Fanny M. Cheung (The Chinese University of Hong Kong)
- Paula Elosua (University of the Basque Country, Spain)
- Kadriye Ercikan (University of British Columbia, Canada)
- Ronald K. Hambleton (University of Massachusetts, USA)
- José Muñiz (University of Oviedo, Spain)
- Kurt Geisinger (University of Nebraska-Lincoln, USA)
- Solange Wechsler (Pontificia Universidade Catolica de Campinas, Brazil)

International Advisory Pannel

- Leandro S. Almeida (University of Minho, Portugal)
- Merry Bullock (IUPsyS)
- Mike Cheung (National University of Singapore)
- Cheryl Foxcroft (Nelson Mandela Metropolitan University, South Africa)
- José Livia Segovia (Peru)
- David Maree (University of Pretoria, South Africa)
- Thomas Oakland (University of Florida, USA)
- Rob Roe (European Federation of Psychologists' Association)
- Yao-Ting Sung (National Taiwan Normal University)
- José Toro (Interamerican Society of Psychology)
- Fons van de Vijver (University of Tilburg, The Netherlands)
- Rafael Vidal (CENEVAL, México)
- Alina von Davier (Educational Testing Service, USA)
- Claudia Zuniga (University of Chile, Chile)

State-Of-Art Lectures

Strong Performers and Successful Reformers in Education

Andreas Schleicher

OECD

e-mail: andreas.schleicher@oecd.org

International comparisons are never easy and they aren't perfect. But the Programme for International Student Assessment (PISA) shows what is possible in education and it helps countries see themselves in the mirror of the educational results and educational opportunities delivered by the world's educational leaders. Even those who claim that the relative standing of countries in PISA mainly reflects social and cultural factors must concede that educational improvement is possible: In mathematics, the students in countries like Brazil, Turkey, Mexico or Tunisia rose from the bottom; Italy, Portugal and the Russian Federation have advanced to the OECD average or close to it; Germany and Poland rose from average to good, and Shanghai and Singapore have moved from good to great. Indeed, of the 65 participating countries, 45 saw improvement in at least one subject area. This is a major tribute to educational efforts made all around the world.

But educational underperformance is still a major challenge. Even in the industrialised world, almost a quarter of 15-year-olds did not even reach Level 2, the baseline in mathematics, where students have to do little more than employ basic algorithms or procedures involving whole numbers. Still, the challenges of school systems are not just about poor kids in poor neighbourhoods, but about many kids in many neighbourhoods. Only 2% of American students reach the highest level of math performance, demonstrating that they can conceptualise, generalise and use math based on their investigations and apply their knowledge in novel contexts. That compares with figures of up to 31% in Shanghai-China. The world economy will pay an ever-rising premium on excellence and a number of countries have shown how the share of top performers in school can be raised significantly, including high performers such as Hong Kong and Korea and low performers such as Italy, Portugal and the Russian Federation.

Of course, raising outcomes is easier said than done and the status quo has many protectors. However, PISA has revealed an encouraging number of features which the world's most successful school systems share. The presentation will shed light on these.

Integrating the Global and the Local in Testing and Assessment

Fanny M. Cheung

The Chinese University of Hong, Hong Kong

e-mail: fmcheung@cuhk.edu.hk

Testing and assessment are increasingly used across cultures. Can we compare people across cultures using the same tests? How do we know what we use to measure people in one culture is valid and relevant in another culture? The etic versus emic aspects of testing at the theoretical, methodological and applied levels have been discussed extensively in the past 50 years. Early efforts in test translation and adaptation assumed universality of constructs and aimed at improving the methodology of equating measures. At the theoretical level, questions have been raised regarding the universality and cross-cultural relevance of the constructs being measured. Attempts have been made to develop indigenous measures to fit the local needs. Given the multicultural interactions within and across geographical boundaries, the dichotomy between the global and the local is no longer a useful paradigm. In a combined emic-etic approach to assessment, both universal and local constructs can be included in a comprehensive framework for cross-cultural comparisons. I will illustrate the combined emic-etic approach with my research program on personality assessment. Adopting this integrative perspective, culture-sensitive approaches can be adopted which identify cultural similarities and respect cultural diversity.

Keynote Addresses

Personality: Individual, Organization, Industry Sector and Country Effects

Dave Bartram

CEB's SHL Talent Measurement Solutions, Thames Ditton, United Kingdom

e-mail: dave.bartram@shl.com

This presentation will review the findings from nearly a decade of research on national and group differences in personality, based on large multinational sets of OPQ32 data. The results will be examined both in terms of the Big Five and at the more detailed 32 primary scale level. The amount of variance in personality scale scores accountable for by countries, by industry sectors and by client organizations will be described and the implications for assessment practice discussed. The relationship of the effects found using forced-choice item types to country level measures of systematic bias in rating scale responding will also be reviewed. What is emerging from this work is a view of how country 'cultural norms' of behaviour may act as constraints on expressed personality, affecting score variance and score level in subtle ways that are relatively small in terms of absolute effect size (typically $d<0.5$) but very robust. These country effects when combined with effect of organizational 'norms' account for around 12% of the variance in individual personality scale scores, with the total being less for some scales and more for others.

Fairness, Validity, and Accessibility: Considerations for Test Design and Development

Linda Cook

Retired, ETS, Lambertville, United States

e-mail: lcook@ETS.org

Test fairness has been a topic of central importance to test developers, test takers, and test users for many decades. Ideas about fair assessments have evolved over time and have stemmed from a variety of disciplines and interests. Recently, fairness and accessibility have come to be thought of as critical factors in the valid interpretations of test scores. That is, unless one can be assured that a test score is a bias free measure of an individual's knowledge, skills, or abilities, the interpretation of the score for a particular use cannot be considered valid. In the United States, the concept of fairness, developed as an important aspect of testing over time and was initially focused on subgroups of the testing population such as males and females and groups defined by race and ethnicity. Changing population demographics in the US and the passage of legislation to protect various subgroups, including individuals with disabilities, has expanded the concept of fairness to include a focus on the accessibility of an assessment for all intended test takers in the test population. The paper begins with a brief exploration of the foundations of fairness in testing in the United States and then provides a more in depth discussion of what is currently done in the US to promote fair and accessible assessments. This discussion includes three basic steps in the process of assessment that lead to test scores that support valid inferences: design, development, and administration. The paper concludes with a section on collecting evidence of the fairness and accessibility of assessments.

Technology: Its Current and Future Effects on Testing

David Foster
Caveon, Mount Pleasant, Utah, United States
e-mail: dfoster@kryteriononline.com

There is no doubt that technologies associated with information presentation, storage, transmission and retrieval have affected the way tests are created, administered and used just as those technologies have affected every other industry in similar ways. Those effects generally include greater convenience, lower costs, greater accessibility, better security, and even improved measurement. But has this technology run its course? Have we seen the extent of the benefits? What new innovations might we expect? Do we use the innovations that exist today or do we hold on to legacy methods and tools? This keynote will present and discuss the relatively brief history of technology and innovation in testing, how we should view innovations and use them today, and what we can expect to see in the years to come.

The Measurement of Learning

John Hattie and Patrick Griffin
University of Melbourne, Carlton, Melbourne, Australia
e-mail: jhattie@unimelb.edu.au

There has been a rich literature, methodology and success in the measurement of many domains of achievement. The same is not so true for the measurement of learning and this Keynote will highlight some of the major learning strategies, overview the meta-analyses on the impact of various strategies, and discuss some of the more recent measures of learning. One major example, the measurement of collaborative problem solving, will be used to illustrate many of the issues, including the use of modern technology, analysing keystrokes leading to dependable scoring models, and rich reporting. The measure of collaborative problem solving will be a focus of the next round of PISA assessment, and opens the way for many other learning processes to be investigated. The approach adopted by PISA will be reviewed in the light of recent research in this area.

Measuring Psychological variables: Current Perspectives and Future Challenges

José Muñiz
University of Oviedo, Oviedo, Spain
e-mail: jmuniz@uniovi.es

The goal is to present the main problems involved in the measurement of psychological variables. Antecedents are outlined, current perspectives analyzed, and possible future developments discussed. Although we will focus on the field of psychology, the arguments presented here can be generalized to the Social and Health Sciences. In order to develop the central idea, the presentation will be organized around seven large axes: Introduction, models of measurement, new technologies, items, reliability, validity, and the use of measurement instruments. In the the origins of psychological measurement and the profound causes that underline the difficulty of measuring psychological variables are analyzed, with special attention to the nature of psychological constructs. Then, the most frequently employed psychometric measurement models are presented, with a clear predominance of the Item Response Theory models. Technology will be the second axis, with special mention of the evolution from paper-and-pencil tests to computerized assessment. The new item formats will be analyzed next. Reliability has evolved towards less global and more analytical forms, such as the Information Function. The same applies to validity, and the new formulations will be underlined. Everything related to the use of tests has become very relevant in the past few years, we comment on the efforts and strategies of national and international organizations to improve the use of tests. The need to combine theoretical and applied approaches to measure psychological constructs will be underlined. Lastly, we reflect on possible future tendencies, underlining that the great impulse of psychological assessment comes from the convergence of two great forces: On the one hand, the new information technologies, particularly the progress in computer science, multimedia, and Internet, and, on the other, the new psychometric models for the treatment and analysis of data.

Behind the scenes: Trials and Tribulations in Building a Personality-Based Development Tool

Eileen Talento-Miller
Graduate Management Admission Council, Reston, United States
e-mail: Talento-Miller@gmac.com

This presentation will provide a case study outlining several critical psychometric challenges that were encountered and resolved in the building of a self-directed, personality-based development tool designed for a specific population. Some of the challenges included identifying the competencies to be assessed and matching learning materials to the competencies. During the presentation, emphasis will be placed on scaling issues which involved defining the reference population and desired score distribution characteristics.

Linking Assessments Internationally with Little or No Data, Few Things in Common and Less Than Motivated Policy Makers: An Empirical Investigation

Jon Twing

Pearson Assessment Centre, Iowa City, IA, United States

e-mail: jon.s.twing@pearson.com

The use of assessments internationally continues to expand as is evident from the increasing popularity of the International Test Commission. With this expansion, practical measurement problems are often exacerbated by a lack of understanding about measurement best practices by some members of ministries around the globe. Concepts and procedures fundamental to sound measurement best practices, like statistical test form linking, are not always used and when used may not be done under optimal assumptions or conditions. This paper will present some of the context and scenarios experienced by the author supporting assessment globally. The paper will explore suboptimal options for linking assessments and will review the research supporting alternative linking schemes in such circumstances. The paper provides examples of research conducted both in the implementation of such designs as well as exploring an original variation of a technique known as "item component equating". The paper concludes with some lessons learned and some options for improving assessment best practice internationally when practitioners are faced with linking assessments with little or no data, few if any commonality and less than motivated policy makers or stakeholders.

Bigger is Often Simpler: Using Shadow Elements in Test Construction

Wim J. van der Linden

CTB/McGraw-Hill Education, Monterey, United States

e-mail: wim_vanderlinden@ctb.com

The shadow-test approach to adaptive testing was primarily introduced to control item selection with respect to any of its content, psychometric, and practical specifications. But the potential applicability of it is much larger. In fact, any test construction problem can be made more effective or efficient by introducing shadow elements in it. We explain the rationale underlying this claim and illustrate its power by a variety of applications.

Invited Symposia

Developments in Europe since 2011: The Work of the EFPA Board of Assessment

Dave Bartram

CEB's SHL Talent Measurement Solutions, Thames Ditton, United Kingdom

e-mail: dave.bartram@shl.com

This session will provide an update on the work of the EFPA Board of Assessment covering the period 2011-2013. The work of the predecessor entity, the Standing Committee on Tests and Testing has been reviewed by Bartram (2011). In 2011 the Standing Committee on Tests and Testing (SCTT), which had a long history of significant developments in the area of test and test user standards in Europe, was replaced by the EFPA Board of Assessment. In the two years it has been in existence, the new Board has been active in completing its agreed work plan. In this session Dave Bartram will provide an overview of the work of the Board from his viewpoint as convenor, with an emphasis on the work done in further developing the test use standards. Jose Muñiz will review the development of the revised Test Review model. Pat Lindley will describe the work of the Test User Accreditation Committee (TUAC). The three presenters will be supported by a Pannel discussant who have all been involved in the work of the Board or TUAC: Sverre Nielsen, Iris Egberink and Lars Michaelsen.

A Review of the EFPA Board of Assessment's Work 2011-2013

Dave Bartram

CEB's SHL Talent Measurement Solutions, Thames Ditton, United Kingdom

e-mail: dave.bartram@shl.com

The review will detail the achievements of the Board including: 1. Completing a survey of European psychologists' attitudes to tests and testing and published the findings. 2. Agreeing a working definition of 'assessment' which is consistent in its scope and form to that used for the ISO 10667 standard for assessment in work and organizational settings (though the EFPA definition covers all areas of assessment); 3. Producing a substantial revision of the EFPA test review model; 4. Updating the EFPA Test Use standards to cover all three domains of practice (work, health and education) each at three levels of competence and producing a document explaining the EFPA Test Use standards and their application; 5. Producing a set of standards for psychological assessment use covering three levels of competence and three domains of practice and producing guidance on the coverage of psychological assessment in the basic EuroPsy for consideration by the EFPA European Awarding Committee.

The Work of the EFPA Test User Accreditation Committee

Patricia Lindley

Newland Park Associates, Wakefield, United Kingdom

e-mail: patlindley@btinternet.com

Test User Standards set out to provide a general European level 'benchmark' against which local national qualification systems can be compared or audited. The present Standards for test use were developed jointly by EFPA and EAWOP. The EuroTest European Certificate in Test Use(EuroTest) is a nationally awarded qualification in test use that has been accredited as meeting the EFPA Standards in Test Use and the quality assurance criteria set out by the EFPA Test User Accrediting Committee (TUAC). TUAC is a Committee set up under EFPA Standing Committee Test and Testing and now the EFPA Board of Assessment with responsibility for

awarding the EuroTest. It delegates the authority to enter the name into the Register and to award the EuroTest in accordance with these Regulations to a National Awarding Committee. The objectives of this work were to be responsible for oversight of the EuroTest and its Regulations and for ensuring that the Registration and award of the EuroTest occurs in accordance with the EFPA, 2) to assess submissions from EFPA member associations of Test User Standards and Qualifications against the EFPA, and to uphold the EFPA Test User Standards. TUAC consists of a Chairperson and at least four other Members appointed for a term of up to four years, once renewable, by the Executive Council of EFPA, where practical the members will be from different countries within the EFPA member states and will represent the relevant professional contexts within which test use is certified and will provide a balance between practitioners and those with specialist expertise. One EFPA member association is now accredited, one awaits decision others actively working towards submission. There is a demand for the EuroTest. It is becoming established and has certificate holders from Europe and beyond.

Revision of the EFPA Test Review Model

José Muñiz

University of Oviedo, Oviedo, Spain

e-mail: jmuniz@uniovi.es

A proper use of tests depends on the quality of the instruments and the expertise of the professionals using them. Therefore, to enhance testing practices it is necessary improving both the tools and the skills of professionals. National and international organizations, such as the International Test Commission (ITC) and the European Federation of Psychologists's Associations (EFPA), have been developing actions and projects for many years to improve tests and testing. The main goal of this presentation is to describe the revised model of the EFPA for the evaluation of the quality of tests. This model aims to provide test users with rigorous information about the theoretical, practical and psychometric characteristics of tests, in order to enhance their use. For the revision of the test review model, an EFPA task force was established, consisting of six European experts from different countries, who worked on the update of the previous European model, adapting it to the recent developments in the field of psychological and educational measurement. The updated EFPA model provides for the exhaustive evaluation of tests. The first part describes test's characteristics, and in the second part, a quantitative and narrative evaluation of the most relevant psychometric characteristics of tests is presented. The EFPA test review model allows a comprehensive analysis of the quality of tests, from a qualitative and quantitative points of view. Finally, the revised model is analyzed in light of recent developments in the field of psychological and educational assessment.

What to Think of Item Response Times?

Paul De Boeck

Ohio State University, Columbus, OH, United States

e-mail: deboeck.2@osu.edu

Item responses in a test always come with a response time. How to interpret these response times? How can one jointly model response times and responses and what can be learned from such models? Are response time data useful to resolve the lingering issue of power and speed? Is there any ability information in response time data? Are fast responses qualitatively different from slow responses? Are we also measuring speed when a test is presented under time

pressure? What can be learned from response times in personality tests? These are important theoretical and practical questions. Perhaps not all these questions will be answered in the symposium, but the audience will see us trying, and, raising the issues may be of interest in itself.

Response Time Effects in Reasoning Considering the Moderating Role of Persons, Items and Item Characteristics

Frank Goldhammer and Johannes Naumann

DIPF - German Institute for International Educational Research, ZIB - Centre for International Student Assessment, Frankfurt am Main, Germany

e-mail: goldhammer@dipf.de

The role of response time in completing an item may have different interpretations. Responding more slowly could be positively related to the probability to obtain a correct response as the item is completed more carefully. However, the association may be negative if working more fluently, and thus faster, reflects higher ability. The objective of this study was to clarify the validity of each assumption for reasoning items. It was assumed that the strength and the direction of the response time effect depend on the items' difficulty as well as persons' reasoning ability. That is, easy items completed by able persons are associated with a negative response time effect and vice versa. Moreover, the goal was to explain item difficulty by the number of rules included in a matrix problem, and, thereby, to determine whether this item characteristic drives the moderating role of item difficulty. A total of 230 persons aged 19 to 40 participated in the study, and completed a computerized version of Raven's Advanced Progressive Matrices test. Results obtained by generalized linear mixed modeling revealed that response time overall had a negative effect. However, this effect was moderated significantly by items and persons. For easy items and able persons the effect was strongly negative, for difficult items and less able persons it was less negative or even positive. The number of rules (i.e., different types of rules and different instances of the same type) involved in a matrix problem proved to explain item difficulty significantly. Most importantly, a positive interaction effect between the number of rules and item response time indicated that the response time effect became less negative with increasing number of rules. These results suggest that time on task has no fixed meaning but depends on individual ability, item difficulty and item characteristics affecting item difficulty.

Fast and Slow Responses in Ability Tests

Minjeong Jeon, Haiqin Chen and Paul De Boeck

The Ohio State University, Columbus, United States

e-mail: jeon.117@osu.edu

Using a two-parameter version of the IRTree model estimated with a new software package in R (flirt) it is shown that fast and slow responses to ability test items could not be differentiated in terms of the ability that is measured, but that the item difficulties do differ in a systematic way. It is a remarkable finding that the same ability is measured while the condition of measurement invariance is not met because of the differences in item difficulty. The difference between fast and slow response difficulties seems related to the gradient and upper asymptote parameters of a model that maps the accuracy level of the responses as a function of the response time. The difficulties of fast responses seem related primarily to the gradient of the curve while the difficulties of slow responses seem related primarily to the upper asymptote of accuracy curve.

A General Item Response Theory Approach to the Analyses of Responses and Response Times on Ability and Personality Tests

Dylan Molenaar

University of Amsterdam, Amsterdan, Netherlands

e-mail: d.molenaar@uva.nl

A general item response theory (IRT) approach to the analyses of responses and response times is outlined. In this approach, separate IRT models are specified for the responses and the response times. These models are subsequently linked by formulating cross relations between them. It is shown which models are appropriate for ability tests and which models are appropriate for personality tests. In addition, it is discussed how popular existing models from the psychometric literature are special cases in this framework. This allows us to compare existing models conceptually and empirically. For instance, it is shown under what circumstances the hierarchical model of van der Linden (2007) and the model by Thissen (1983) coincide. Extensions of the traditional models are proposed motivated by practical problems. In addition, some real data analyses are presented.

Considerations and Guidelines on the Fair Assessment of Linguistically Diverse Populations

Alina von Davier[1] and Paula Elosua[2]

[1]Educational Testing Service, Princeton, NJ,United States; [2]University of Basque Country, San Sebastian, Spain

e-mail: avondavier@ets.org

In this symposium, we will present various considerations associated with increasing fairness in the assessment of linguistically diverse populations occurring when assessing linguistic complex comunities within a country, or when tests are repurposed and are administered across countries either in their original language or in other languages. Within this context, the first two presentations provide broadly applicable guidelines associated with assessing linguistically diverse populations. The discussion will focus on various aspects of test and item development, test design, test fairness and test validation associated with developing and administering assessments in these contexts. These contexts often differ from those in which tests are developed from the beginning to be administered internationally to diverse linguistic populations. The objective of these presentations is to present a view of the challenges associated with developing and administering tests to linguistically diverse populations and to provide recommendations for improved test design and development practices. The next three presentations will provide examples of the kinds of considerations that need to be implemented in particular types of assessments contexts. Specifically, in the third presentation, psychometric and test validation considerations given in the context of using personality assessments (e.g., the NEO-FFI 3) with culturally and linguistically diverse populations are discussed. In the fourth presentation, a focus is given to complexities (e.g., investigating the impact of test language and cultural status on vocabulary and working memory performance) arising in the context of administering educational achievement tests (e.g., various vocabulary assessments). The fifth presentation discusses psychometric and psycho- and sociolinguistic considerations within the context of mathematics achievement in linguistically diverse contexts in the Basque Autonomous Community. These presentations will serve to advance research on the assessment of linguistically diverse populations by exemplifying the various challenges associated with their

assessment and to provide suggestions for improved practices that apply generally or within particular contexts.

Effects of Language on Math Achievement Performance in the Bilingual Context of the Basque Autonomous Community

Paula Elosua[1] and Paul De Boeck[2]

[1]University of Basque Country, San Sebastian, Spain; [2]Ohio State University, Columbus, OH, United States

e-mail: paula.elosua@ehu.es

The aim of this paper was to study math achievement performance in the bilingual Basque Autonomous Community (BAC) from four perspectives: linguistic, psycholinguistic, sociolinguistic and psychometrics. Using a multilevel logistic modeling approach the effects of factors that may play a role in a linguistic diversity context was investigated. The data are from a large sample (N=15,401) of students grouped according to their home language (Spanish or Basque) and test language (Spanish or Basque). The results show that not only psychometric measurement equivalence is important. Also other aspects such as the sociolinguistic factors and the linguistic distance between languages seem important in order to draw valid from an educational assessment program.

Cross-Linguistic and Cross-Cultural Effects on Verbal Working Memory and Vocabulary: Testing Language Minority Children with an Immigrant Background

Pascale Engel De Abreu[1], Martine Baldassi[2], Marina Puglisi[3] and Débora Befi-Lopes[3]

[1]University of Luxembourg, Walferdange, Luxembourg; [2]Columbia University, New York, United States; [3]University of Sao Paulo, Sao Paulo, Brazil

e-mail: pascaleengel@gmail.com

The study explored the impact of test language and cultural status on vocabulary and working memory performance in multilingual language minority children. Twenty 7-year-old Portuguese-speaking immigrant children living in Luxembourg completed several assessments of first- and second-language vocabulary (comprehension and production), executive-loaded working memory (counting recall and backward digit recall), and verbal short-term memory (digit recall and nonword repetition). Cross-linguistic task performance was compared within individuals. The language minority children were also compared with multilingual language majority children from Luxembourg and Portuguese-speaking monolinguals from Brazil without an immigrant background matched on age, sex, socioeconomic status, and nonverbal reasoning. Results showed that (a) verbal working memory measures involving numerical memoranda were relatively independent of test language and cultural status; (b) language status had an impact on the repetition of high- but not on low-wordlike L2 nonwords; (c) large cross-linguistic and cross-cultural effects emerged for productive vocabulary; (d) cross-cultural effects were less pronounced for vocabulary comprehension with no differences between groups if only L1-words relevant to the home context were considered. The study indicates that linguistic and cognitive assessments for language minority children require careful choice among measures to ensure valid results. Implications for testing culturally and linguistically diverse children are discussed.

Psychometric Properties of a Measure of Personality as a Function of Language Literacy and Test-Taking Motivation in an Ethnically Diverse Sample

Dragos Iliescu[1] and Ion Andrei[2]

[1]SNSPA University, Bucharest, Romania; [2]Bucharest University, Bucharest, Romania

e-mail: dragos.iliescu@testcentral.ro

A number of psychometric characteristics of the NEO-FFI-3 (McCrae & Costa, 2010) were investigated for the Romanian language version of the test, in an ethnically heterogeneous sample of highschool students aged 16-19 years. Specifically, indicators of reliability (internal consistency and item-total correlations) and of validity (construct validity and various forms of invariance) were investigated. The sample contains almost 600 children living in Romania in Romanian and non-Romanian (Gipsy, Hungarian and Russian) communities. Supplementary measures of Romanian literacy and of motivation were collected for all children. The school grades for Romanian language were collected from school records. The test administrator also assessed language proficiency based on a short 15-minutes pre-assessment interview. All children were further assessed with a 3-item pre-test measure of test motivation modeled after the pretests scale used by Eklof (2006a) in investigating the motivation of Swedish children for the TIMSS exam. Also, all children were assessed with the Student Opinion Scale (Sundre & Moore, 2002) as a post-test. The analysis shows that Romanian literacy and pre-test motivation influence both the reliability and validity of the NEO-FFI-3 for children who come from backgrounds which are ethnically different from the mainstream culture, but not for Romanian children.

A Validity Argument to Develop and Use Exported Assessments Fairly*

Maria Elena Oliveri, Rene Lawless and John W. Young

Educational Testing Service, Princeton, NJ, United States

e-mail: moliveri@ets.org

A trend in educational assessment has led to increased demands for administering assessments to more diverse audiences. Their fair use requires score comparability across populations to minimize construct-irrelevant variance (e.g., lack of familiarity with the test language, item types, or testing conditions) so as to not confound the various populations' ability in demonstrating knowledge of the assessed construct. We address this issue by presenting a validity argument for fair assessment development based on the notion that fairness ensures score consistency across diverse populations (Kane, 2013). Our framework considers the impact of this phenomenon on six dimensions (domain definition, evaluation, generalization, explanation, extrapolation, and utilization) relating how different aspects of an assessment may impact test fairness for (linguistically) diverse examinee populations. The purpose is to identify ways to evaluate, quantify, and minimize sources of potential construct-irrelevant variance and increase test fairness. We incorporate evidence centered design methodology (Mislevy, Steinberg, & Almond, 1999; Mislevy & Haertel, 2006) to identify the relevant evidence necessary to assess the construct, map out the evidence necessary to assess examinees' performances to support the inferences made regarding their competencies, and consider the factors that may potentially cause construct-irrelevant variance. We exemplify components of the framework using contextualized examples from: 1) the Graduate Records Examination®, developed for admissions decisions into U.S. graduate schools, which is being considered for admissions into Asian graduate schools using English as the language of instruction; and 2) el examen de Admision a Estudios de Posgrado®, developed for use in Puerto Rico, which is administered to other Spanish-speaking countries. Both uses raise fairness challenges. We identify

generalizability challenges and provide recommendations to increase test fairness for exported assessments. We will provide a systematic and objective framework explicating the steps necessary to develop a validity argument and improve test development/design practices for exported assessments.

International Test Commission Guidelines for Assessing Linguistic Minorities

Alina von Davier, Maria Elena Oliveri and René Lawless
Educational Testing Service, Princeton, NJ, United States
e-mail: avondavier@ets.org

Several complexities arise when examinees who speak a language other than the language of the test are assessed. These may arise in countries with high levels of immigration or that have more than one official language or when a test is used in multiple countries. As an international organization devoted to promoting fair and valid assessments around the globe, members of the International Test Commission are developing a set of guidelines to raise awareness about the problems in measuring the knowledge, skills, and abilities of linguistic minorities and to provide guidance to test developers, test administrators and those who interpret test scores for such test-takers. In this presentation, we will illustrate the guidelines being developed to address these complexities from several perspectives: test and item development, test design and psychometric analyses, test validation, scoring and score interpretation, testing accommodations, cross-cultural dimensions in assessments, quality control, among other special features of psychological and educational assessments. A discussion synthesizing these perspectives will be provided.

Assessments of Reading Literacy in Different Languages

Kadriye Ercikan
University of British Columbia, Vancouver, Canada
e-mail: kadriye.ercikan@ubc.ca

Development of reading literacy is central to schooling and education broadly. Measurement of reading literacy is therefore key to examining and monitoring learning outcomes throughout the world. International assessments play an important role in guiding policy and practice in the participating countries in addition to providing an international perspective on educational practices and policies. Two international assessments the Progress in International Reading Literacy Study (PIRLS) and the Programme for International Student Assessment (PISA) include assessments of reading and language literacy. Both international assessments face the challenge of measuring reading literacy in multiple languages. In this symposium, the presenters will discuss research that addresses challenges in developing measurements and comparability of measurements of reading literacy across different languages. The first presentation by Dr. Serge Lacroix will address the development of an achievement test in French for children studying in a linguistic minority setting, and the comparability of tests developed in English then translated and adapted for a French population in Canada. Following this presentation, Dr. Paula Elousa will discuss comparability of reading assessments in Spain with many linguistic groups focusing on Basque, Catalan, Galician and Spanish. The last two presentations will focus on comparability of reading measurement across languages in PIRLS administered to examinees in 48 countries in 2011. The presentation by Dr. Arim (co-authors Juliette Lyons-Thomas) will examine measurement comparability between English and French versions of PIRLS administered in

Canada, United States and France. The last presentation by Dr. Kadriye Ercikan (co-author Eugene Gonzalez) will examine differences in country specific score scales and the international scale used in creating scores in PIRLS. Together, these studies will contribute to research on factors that affect comparability of reading literacy measurements and degree of comparability of measurement in a very high profile international assessment. The symposium will be chaired by Eugenio Gonzalez.

Measuring Reading Skills in Various Languages Using the Same Test

Serge Lacroix

University of British Columbia, Vancouver, Canada

e-mail: auguston@shaw.ca

Canada is a bilingual country where both English and French are taught, starting in elementary years and often all the way to the secondary school graduation. Because students are assessed in either or often both languages, the issue of comparability of scores is an important one. Tests used, more specifically those that assessed reading skills, are explored as part of this symposium paper. Through this paper, the author will look more at two issues: a) the development of an achievement test in French for children studying in a linguistic minority setting and b) the comparability of tests developed in English then translated and adapted for a French population. Developing a test for a linguistic minority population comes with direct challenges associated with the bilingual population for which the test is created. As an example, reading tests developers have to avoid using words that are too close in pronunciation or meaning, including "faux-amis" (false cognates) to avoid measuring the wrong construct or a construct that is different from the one intended. This supplemental step makes them more aware of the issues relating to item development. Researchers that translate and adapt reading tests are confronted with different challenges as they have to work within the confine of an already existing version of the test while also wanting to measure similar construct. The primary purpose of this presentation will be study to address these issues and illustrate how they impact the comparability of the test scores obtain by various populations. Using differential item functioning (item response theory), we will look at comparisons of item functioning and offer suggestions as what could explain it, using the information provided above.

Comparability between English and French Versions of the PIRLS 2011 Reader Test Results from Canada, United States, and France

Rubab Arim[1], Juliette Lyons-Thomas[2] and Kadriye Ercikan[2]

[1]Ottawa Hospital Research Institute, Ottawa, Canada; [2]University of British Columbia, Vancouver, Canada

e-mail: rarim@ohri.ca

One of the key outcomes of international assessments is a summary report that provides a comparison of international results. There has been a plethora of research showing that valid comparisons across countries require comparability of scores across countries. Additionally, there are many factors, such as curricular, cultural, and language differences between countries, which may affect this comparability. In international assessments, one obvious source of differences between scores from different country administrations is the different language versions of the test forms. Previous research emphasized the importance of examining comparability of test forms in different languages at the item level using differential item functioning (DIF) analysis methods and at the test level using statistical analyses that compare test data structure, such as factor analysis. The primary objective of this study was to examine the degree of measurement comparability between English and French versions of the Progress

in International Reading Literacy Study (PIRLS) 2011 Reader test results from Canada, United States, and France. Measurement invariance will be assessed for the same language version of tests between countries (e.g., Canada French and France) as well as between the two language versions of tests within (e.g., Canada English and French) and between countries (e.g., United States and France) using DIF analyses and multi-group confirmatory factor analyses. The impact of the differences on the score scale comparability will also be examined by comparing the test characteristic curves. The findings will be discussed in light of local and global challenges for best practices in international assessments.

A Framework for Examining Accuracy of a Single Scale for Multiple Language Versions of Assessments: The Special Case of International Assessments of Learning Outcomes

Kadriye Ercikan[1] and Eugenio Gonzalez[2]

[1]University of British Columbia, Vancouver, Canada; [2]Educational testing Service, Princeton, NJ USA, United States

e-mail: kadriye.ercikan@ubc.ca

There has been a significant amount of research on measurement comparability and score scale comparability of multiple language versions of tests administered to examinees from different language and cultural backgrounds. Professional standards informed by this research require that if scores from different language versions are intended to be comparable, evidence of comparability should be provided. However, there is no specific guidelines about what sufficient evidence is and when separate score scales for different language versions should be provided. In this article we provide a framework that proposes a specific methodology for determining accuracy of a single score scale for multiple language versions of assessments. Data from the Progress in International Reading Literacy Study is used to elaborate on the proposed framework and the accuracy of a single international scale for individual countries participating in the assessment is examined.

PISA 2009 Reading Comprehension Tests in Spain. Differences among Languages

Paula Elosua and Josu Mujika

University of Basque Country, San Sebastian, Spain

e-mail: paula.elosua@ehu.es

The PISA project provides the basis for studying curriculum design and for comparing factors associated with school effectiveness. These studies are only valid if the different language versions are equivalent to each other. In Spain, the application of PISA in autonomous regions with their own languages means that equivalency must also be extended to the Spanish, Galician, Catalan and Basque versions of the test. The aim of this work was to analyse the equivalence among the four language versions of the Reading Comprehension Test (PISA 2009) and to study differences in performance After defining the testlet as the unit of analysis, equivalence among the language versions was analysed using two invariance testing procedures: multiple-group mean and covariance structure analyses for ordinal data and ordinal logistic regression. The procedures yielded concordant results supporting metric equivalence across all four language versions: Spanish, Basque, Galician and Catalan. The equivalence supports the estimated reading literacy score comparability among the language versions used in Spain.

Innovative Solutions to Changing Measurement Priorities

Ronald Hambleton
University of Massachusetts, Amherst, MA, United States
e-mail: rkh@educ.umass.edu

This is a period of significant change in measurement. Much of this change is driven by different expectations of test users and changes in education and the workplace. Many users expect more understanding of what test performance represents about a person's ability than is provided in a single score. As technology, outsourcing, and other changes in the workplace eliminate work, jobs that were formerly routine now require higher order thinking skills and licensure and certification tests need to change to reflect those new requirements. Demands for more realistic assessment scenarios, efficient testing, and convenience for test takers also affect test design, development, and delivery. The measurement profession is responding to the demands in a variety of ways including several forms of computer-based testing, remote proctoring, assessment engineering, evidence centered design, performance testing, gamification, and more. This symposium will discuss how three organizations, The American Institute of Certified Public Accountants (AICPA), the Medical Council of Canada (MCC), and CEB have innovated to respond to changes in their environments and customer needs. A fourth presentation from our discussant will discuss implications of these innovations for other licensure, employment, and education measurement programs. Craig Mills, AICPA, will discuss the application of evidence-centered design to the design a practice analysis and the development of item formats measuring higher-order skills. André-Philippe Boulais, MCC, will discuss the MCC's research in the areas of automated item generation and automated scoring of clinical decision making tasks. Eugene Burke, SHL, will discuss gamification for situational judgment tests and the development of a self-assessment to predict safety and customer service outcomes. Ronald Hambleton, UMASS, Amherst, will summarize how the work of these organizations related to similar work elsewhere in the profession and discuss the implications of the work of these three organizations for other testing programs.

Using Assessment Engineering to Guide a Practice Analysis and Develop Innovative Items

Craig Mills
CTB/McGraw Hill, Monterey, CA, United States
e-mail: craig.mills@ctb.com

The Uniform Certified Public Accountant (CPA) Examination is one of the requirements for licensure as a US CPA. Traditionally, the focus of test development and construction has been to ensure that content requirements are met. Less emphasis has been placed on identifying and describing the underlying cognitive skills being tested. Thus, the tests have predominantly measured knowledge and understanding of technical accounting knowledge. Although skills are assessed through the "application of the body of knowledge" skills assessment remains focused on technical skills (identifying relevant citations in standards, etc.) While the current exam has served the profession well, several trends are changing the expectations for entry-level CPAs. Applications of technology, outsourcing, and use of paraprofessionals for many routine tasks have eliminated much of the work previously assigned to new CPAs. As a result, new CPAs are expected to perform at higher cognitive levels than was previously required (e.g. to review rather than prepare documents). This has placed additional importance on skills such as communication and analytical thinking in additional to technical knowledge. In fact, a recent analysis suggested that written communication and analytical thinking are more important than technical knowledge

with quantitative problem solving skills and oral communication following closely behind. The CPA Exam will need to evolve to reflect the changes needed at entry. The AICPA is applying evidence centered design concepts to the design of its new practice analysis and the development of task models and item templates to be used to determine the content and skills to be included in a revised examination. This paper will summarize the current work being conducted to support the practice analysis and will provide examples of item templates derived from evidence centered approaches at the AICPA.

Breaking the Normal Rules of Psychometrics to Address Client Needs in the Employment Space

Eugene Burke

CEB SHL Talent Management Solutions, Thames Ditton, United Kingdom

e-mail: eugene.burke@shl.com

Increased technology in assessment brings opportunities and challenges. The opportunity via Internet testing is the scalability and reach that client organizations can draw upon in competing for talent. There are challenges as well. One is pressure on the assessment window – the time required to complete an assessment – that comes from concerns about the candidate experience when candidates may be employees or customers. Another is the desire for more immersive and, for Gen Y, more gamification of assessments. Addressing these challenges requires a reframe of assessment design as well as a rethink of existing paradigms. An example of addressing the assessment window is described through how a 7 minute self-assessment was developed to target key behaviors for operational roles that shows strong criterion and construct validity. Data on both facets of validity are provided. With more immersive assessments comes the challenge of construct validity and, below the surface of the presentation layer that the candidate experiences, the question of what exactly does the simulation measure? How this question can be addressed is described through a reframing of the development process for situational judgment tests (SJTs) using a behavioral framework. Evidence for criterion and construct validity are shared. Both examples show that existing frames for assessment development may need to be reconsidered and some rules broken to meet the opportunities and the challenges presented by hi-tech assessments.

Innovations in Medical Licensure in Canada

André-Philippe Boulais

Medical Council of Canada, Ottawa, Canada

e-mail: aboulais@mcc.ca

The examinations required for medical licensure in Canada include a nationally standardized assessment of knowledge, competencies, skills, and attitudes required for safe practice. The examination includes two parts: Part I is a computer-administered examination composed of an MCQ adaptive test and a decision-making component which elicits written responses. Part II is an objective structured clinical examination (OSCE) where live interactions with standardized patients are scored by experienced physicians. The MCC has adopted an ambitious validity research agenda in support of its strategic goals. Included in that agenda is a desire to streamline the item development process. The feasibility and value of using cognitive models for item authoring have been studied. A feasible solution has been developed with Dr. Mark Gierl and Dr. Hollis Lai at the University of Alberta: Automated Item Generation (AIG). This research may benefit the MCC as it explores methods to efficiently produce large numbers of high quality test items in response to a demand for more frequent and flexible administrations of these

examinations. This presentation will describe the development and quality of test questions generated using AIG. Advances in machine learning and natural language processing are also affecting high-stakes, large-scale assessments. At MCC, the viability of implementing an automated scoring system for constructed response items was studied. In this part of the presentation we will outline agreement rates obtained between machine-scored and physician-scored responses using open-source natural language processing in the context of a medical licensing examination offered in English and in French. The viability of this automated scoring system was assessed using agreement rates and Kappa coefficients. The suggested framework is adaptable and has provided strong evidence to support the use of automated scoring for a high-stakes, multilingual medical licensing examination.

International Career Adaptability Project: Models and Measures

Frederick Leong
Michigan State University, East Lansing, MI, United States
e-mail: fleong@msu.edu

In 2008, the International Career Adaptability Project first gathered in Berlin, Germany, with the overarching goal of creating and developing a measure of career adaptability (Leong & Walsh, 2012). The collaborative defined career adaptability as a social competency which serves as a resources for individuals as they cope with current and future career tasks, transitions and traumas (Savickas, 2005). Career adaptability is defined as an individual's readiness and resources for handling current and anticipated tasks, transitions, and traumas in their occupational roles that alter their social integration (Savickas, 2005). It was from this international collaborative that the Career Adapt-abilities Scale (CAAS) was formed (Savickas & Porfeli, 2012). The first stage of this program of research was focused on the internal validity of the scale. In a special issue of the Journal of Vocational Behavior, researchers from 13 countries reported primarily on the structural or internal validity of the scale (Leong & Walsh, 2012). The papers on this symposium will focus on continuing international studies of the career adaptability construct.

Assessing the External Validity of the Career Adapt-Abilities Scale (CAAS)

Frederick Leong
Michigan State University, East Lansing, MI, United States
e-mail: fleong@msu.edu

In the present study, we move to the next stage of the program of research by assessing the external validity of the scale among a large group of U.S. college students who had part-time work experience. Specifically, we extend Savickas & Porfeli's (2012) study by examining the degree to which career adaptability would be predictive of theoretical related outcomes. In terms of work related outcomes, we selected job satisfaction and work engagement as the criterion variables. Moving to broader criteria, we also assessed the relationships between career adaptability and subjective well-being and coping strategies. As predicted, CAAS positively predicted these set of measures which provides preliminary support for the external validity of the scale. An additional goal of our study was to examine the differential validity of the CAAS with the construct of general adaptability which was operationalized by the I-DAPT measure (Ployhart & Bliese, 2006). We predicted and found that the CAAS was not significantly correlated with the I-

DAPT measure. The implications of these findings and some possible directions for future research are also discussed in this paper.

The Influence of Career Adaptability in the Process of Entering and Remaining in the Working World

Marcelo Afonso Ribeiro

Institute of Psychology - University of Sao Paulo, São Paulo-SP, Brazil

e-mail: marcelopsi@usp.br

Today's working world has been configured in a flexible, heterogeneous and complex way, what has been breaking the previously stability established and has been letting workers with a lack of social and work references, living on flexicurity as the only current possibility of certainty in the workplace. Due to this situation, career adaptability has turned into a basic contemporary competence for the entering and remaining at the working world and for the construction of the work project. Based on the Life Design theoretical model, the proposal presented here aimed to understand the ways that a group of urban workers in Sao Paulo (Brazil) has dealt with the issues of entering and remaining at the working world through a content analysis of their narratives. The main results showed that the career adaptability has been developed in different ways in the study group, and both the construction of a work project based on flexibility or on stability are different ways of developing career adaptability and can generate success or failure in the entering and permanence the working world. As a conclusion, the heterogeneity and the complexity of the current working world have required assumptions that ought to help in the understanding of it, and, at the same time, it might give support in the construction of analysis categories of the psychosocial phenomena of the working world, which can be operationalized into measures for assessment.

Career Adaptability, Hope and Life Satisfaction in Workers with Intellectual Disability

Laura Nota, Salvatore Soresi, Sara Santilli and Maria Cristina Ginevra

University of Padova, Padova, Italy

e-mail: laura.nota@unipd.it

As known, the socio-economic situation impacting most of Europe is characterized by recession, austerity measures, and funding cuts in social services and public assistance. The most immediate real economy consequences of these changes have been those of business closure and job loss (O'Reilly et al., 2011). The unpredictable and unstable current work market is impacting in particular at-risk workers, such as individuals with disability. Based on Life Design approach, the present study is focusing on two variables, hope and career adaptability, relevant to coping with the current work context and their role in affecting life satisfaction. 120 (60 women and 60 men) adult workers with mild intellectual disability are involved. We ask them to complete the Career Adapt-abilities Scale (Soresi, Nota & Ferrari, 2012), Adult Hope Scale (AHS, Snyder et al., 1991). and The Satisfaction with Life Scale (SWL, Diener et al. 1985). In this work the relationships between relationships between career adaptability, hope and life satisfaction will be explored. These results have important implications for practice and underscore the need to support workers with disability in their life design process.

Forced-Choice Measurement: New Developments

Alberto Maydeu-Olivares[1] and Anna Brown[2]

[1]University of Barcelona, Spain; [2]University of Kent, Canterbury, United Kingdom

e-mail: a.a.brown@kent.ac.uk

Despite their unique value in predicting important life outcomes, personality and other non-cognitive attributes are among the most challenging to measure. In employment testing contexts, self-reports are compromised by applicants and employees "faking good" (Birkeland et al., 2006). Responses are also open to unmotivated distortions, such as idiosyncratic uses of rating categories or agreeing with questions overall (e.g. van Herk, Poortinga & Verhallen, 2004). To reduce such distortions, test items may be presented in so-called 'forced-choice' formats. Respondents may be asked to rank-order a number of items, or distribute a fixed number of points between several items – therefore they are forced to make a choice. Until recently, basic classical scoring methods were applied to such formats, leading to scores relative to the person's mean (ipsative scores). Recent advances in estimation methods enabled rapid development of item response models to infer proper measurement from forced-choice data. This symposium will present latest developments in designing and scoring forced-choice questionnaires. The contributions focus on optimal questionnaire design under the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011); Computerized Adaptive Testing (CAT) using the Multi-Unidimensional Pairwise Preference Model (Stark, Chernyshenko & Drasgow, 2005); and a study of resistance of the forced-choice formats to motivated response distortions.

Computerized Adaptive Personality Testing: Methods to Meet the Challenges of High Stakes Uses

Stephen Stark[1], Oleksandr Chernyshenko[2], Fritz Drasgow[3] and Christopher Nye[4]

[1]University of South Florida, Tampa, FL, United States; [2]Nanyang Technological University, Nanyang Business School, Singapore; [3]University of Illinois, Department of Psychology, Champaign, IL, United States; [4]Michigan State University, Department of Psychology, East Lansing, MI, United States

e-mail: sestark@usf.edu

Over the last 15 years computerized adaptive testing (CAT) in educational and organizational settings has increased dramatically. Although cognitive ability and knowledge tests still constitute the majority of applications, noncognitive CAT is expanding rapidly due to evidence that personality (Paunonen & Jackson, 2000) and vocational interest (Nye et al., 2012) constructs, for example, predict retention and performance outcomes. Moreover, in contrast to substantial evidence indicating the importance of broad cognitive factors, research suggests there is much to be gained in noncognitive measurement by focusing on narrow factors due to their relatively low intercorrelations and relationships that vary in magnitude and sign with respect to outcomes. Narrow factor representations of the Big Five personality taxonomy typically involve three to six facets per broad factor. To develop detailed profiles of examinees for selection and classification thus requires long testing sessions in the absence of efficient methods for choosing items. To complicate matters, the potential for faking makes it imperative to consider formats other than Likert-type items and to utilize methods for thwarting and detecting aberrant responding. We will present research on adaptive personality testing with a multidimensional pairwise preference format, which was chosen to reduce faking while keeping the response process simple for participants. In field applications, tests are constructed and scored using the multi-unidimensional pairwise preference item response theory model (MUPP; Stark, 2002; Stark et al., 2005; 2012), and answer patterns are screened for aberrance using response latencies, Markov transition, and standardized log likelihood indices in real time, so that

decision makers can flag suspect patterns for additional screening and researchers can later compare psychometric properties and test validities for "normal" and "aberrant" examinee groups. Our presentation will discuss principles of test construction, scoring, and aberrance detection, and review simulation and empirical evidence concerning CAT validities in various settings.

Data-Driven Development of Forced-Choice Measures: A Pilot Study

Yin Lin[1], Ilke Inceoglu[2], Mathijs Affourtit[1] and Anna Brown[3]

[1]CEB, Thames Ditton, Surrey, United Kingdom; [2]Surrey Business School, University of Surrey, Guildford, United Kingdom; [3]University of Kent, Canterbury, Kent, United Kingdom

e-mail: yin.lin@shl.com

The forced-choice response format has recently gained more popularity following resolution of the ipsativity problem through Thurstonian item response theory (IRT) modelling (Brown & Maydeu-Olivares, 2011, 2013). With its superior resistance to response distortions (Cheung & Chan, 2002), forced-choice formats are especially desirable in high-stake assessment situations (Christiansen et al., 2005). While Thurstonian IRT modelling can improve existing forced-choice instruments (Brown & Bartram, 2009), developing forced-choice measures from scratch typically relies predominantly on subject matter expert (SME) judgements to balance content and optimise measurement. In this study, a data-driven process was piloted to develop a forced-choice interest questionnaire based on Holland's RIASEC model (Holland, 1985). An initial item pool was developed and trialled in single-stimulus format. Data analysis selected 60 items with adequate content coverage of the 6 RIASEC constructs and good confirmatory factor model fit. Samejima's (1969) graded response model was then fitted to obtain item parameters, which were used as approximate Thurstonian IRT parameters for information calculations. The items were then assembled into 20 triplets, optimising IRT information with a target standard error of 0.5 theta or below across all scales for the average person. Each triplet and the entire questionnaire were checked against quantitative content criteria and by SMEs. A few iterations were needed to develop three satisfactory forms. Following simulations to check score recovery, the best performing form was selected for trialling. A sample of 3137 respondents completed the forced-choice questionnaire. For five of the six scales, the standard error target was achieved by between 71.1% and 99.7% (mean 86.1%) of the sample. However, the Social scale missed the target completely, giving a mean conditional standard error of 0.55 theta across the sample. Further analysis will identify factors which may have contributed to successes and pitfalls. Lessons learned and recommendations will be discussed.

Optimal Forced-Choice Measurement for Workplace Assessments

Safir Yousfi[1] and Anna Brown[2]

[1]German Federal Employment Agency, Nuremberg, Germany; [2]University of Kent, Canterbury, Kent, United Kingdom

e-mail: safir.yousfi@arbeitsagentur.de

Forced-choice questionnaires have been shown to improve personality assessment by reducing bias commonly affecting single-stimulus items (e.g. idiosyncratic use of rating scales, acquiescence, halo effects, socially desirable responding). Recent progress in psychometric modelling has overcome the limitations of ipsative scoring, to date the most prevalent method of analysing forced-choice data (Brown & Maydeu-Olivares, 2013). Recommendations for assembling forced-choice questionnaires based on this recent approach are available (Brown & Maydeu-Olivares, 2011). Nevertheless, designing a forced-choice questionnaire remains a

challenge for test developers as numerous options for assigning items to forced-choice blocks have to be evaluated with respect to many different criteria. Optimal test design with mixed integer programming offers a powerful tool to address such complex tasks of test assembly. The present study demonstrates how the of forced-choice questionnaire design can be expressed in formal equations that can be passed as input to open source solvers for mixed integer programming problems. The algorithmic solution of the test assembly problem is compared to two types of test forms: (1) forms assembled at random, without prior knowledge of items parameters, and satisfying basic design constraints only (e.g. number of items per trait); and (2) forms developed by experts carefully considering available recommendations for forced choice questionnaire assembly. The three methods are illustrated with an empirical example of a questionnaire of personal competencies, and are compared for precision of measurement they provide. Implications for the process of developing forced-choice questionnaires are outlined.

Validating Educational and Psychological Assessments: Standards, Applications, and Current Viewpoints

Stephen G. Sireci

University of Massachusetts Amherst, MA, United States

e-mail: sireci@acad.umass.edu

Validating Educational and Psychological Assessments: Standards, Applications, and Current Viewpoints Symposium Summary for the 2014 Conference of the International Test Commission The Standards for Educational and Psychological Testing, developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, represent over 50 years of research and consensus opinion regarding the concept of validity and test validation. The Standards specify five sources of validity evidence that can be used to support the use of a test for a particular purpose. This symposium will feature 6 paper presentations, one on each source of validity evidence, and one on how the five sources impact various testing applications. The presentations will be discussed by a distinguished leader in assessment and validation. The lineup for the session follows. Chair: Stephen G. Sireci, University of Massachusetts Amherst, USA Presentation 1: "Validity Evidence Based on Test Content" Stephen G. Sireci and Molly Faulkner-Bond, University of Massachusetts, USA Presentation 2: "Validity Evidence Based on Relations to Other Variables" Carmel Oren, Avi Allalouf, and Yoav Cohen, National Institute for Testing and Evaluation, Israel Presentation 3: "Validity Evidence Based on Internal Structure" Joseph Rios and Craig S. Wells, University of Massachusetts Amherst, USA Presentation 4: "Validity Evidence Based on Testing Consequences" Suzanne Lane, University of Pittsburgh, USA Presentation 5: "Validity Evidence Based on Response Processes" José-Luis Padilla and Isabel Benitez, University of Granada, Spain Presentation 6: "Developing Sources of Validation Evidence across Assessment Settings" Wayne Camara, ACT, USA Discussant: José Muñiz, Universidad de Oviedo, Spain.

Demonstrating the Validity of Three General Scores of PET in Predicting Higher Education Achievement in Israel

Carmel Oren

NITE, Jerusalem, Israel

e-mail: carmel@nite.org.il

The Psychometric Entrance Test (PET), used for admission to higher education in Israel together with the Matriculation (Bagrut), had in the past one general score in which the weights for its domains: Verbal, Quantitative and English, were 2:2:1, respectively. In 2011, two additional total scores were introduced, with different weights for the Verbal and the Quantitative domains. This study compares the predictive validity of the three general scores of PET, and demonstrates validity in terms of utility. Sample was composed by 100,863 freshmen students of all Israeli universities over the classes of 2005-2009. Regression weights and correlations of the predictors with FYGPA were computed. Simulations based on these results supplied the utility estimates. On average, PET is slightly more predictive than the Bagrut; using them both yields a better tool than either of them alone. Assigning differential weights to the components in the respective schools further improves the validity. The of the new general scores of PET is validated by gathering and analyzing evidence based on relations of test scores to other variables. The utility of using the test can be demonstrated in ways different from correlations.

Validity Evidence Based on Response Processes in Cross-Lingual and Cultural Testing

Jose-Luis Padilla and Isabel Benitez

University of Granada, Granada, Spain

e-mail: jpadilla@ugr.es

Validity evidence based on response processes was first introduced explicitly as a source of validity evidence in the latest edition of the Standards for Educational and Psychological Testing (AERA et al., 1999). However, there are no clear indications and suggestions on contents, methods or the scope of this source of validity evidence. The aim of the presentation is (1) to determine when it is critical to have evidence based on the response process to support the use of the test for cross-lingual or cultural comparisons; and (2) present methods available for conducting validation studies on response processes in cross-lingual or cultural settings, with special attention to the cognitive interview method. Together with a brief systematic literature review, theoretical and practical argument will be discussed. In addition, examples of validation studies conducted to obtain validity evidence based on response processes will be presented. Arguments for determining when validity evidence based on response processes is critical for supporting the use of the test or questionnaire in cross-lingual and cultural setting, along with indications on how to conduct a validation study aimed at obtaining validity evidence based on response processes. Validity evidence based on response processes play a key role in supporting the use of test and questionnaires in cross-lingual and cultural settings. Qualitative methods like cognitive interviewing are very helpful to obtain such kind of validity evidence.

Validity Evidence Based on Internal Structure

Joseph Rios and Craig Wells
University of Massachusetts Amherst, MA, United States;
e-mail: jarios@educ.umass.edu

The concept of validity has evolved dramatically from Guilford's (1946) definition that "…a test is valid for anything with which it correlates" (p. 429). The current conception of validity entails a unitary framework for which there are five sources of evidence that are based on: 1) test content, 2) internal structure, 3) relations to other variables, 4) response processes, and 5) consequences of testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The purpose of the present paper is to describe basic methods for evaluating the internal structure of a test with respect to dimensionality, measurement invariance, and reliability. In describing dimensionality assessment, particular attention is given to explaining both the bifactor model, as well as model selection techniques that go beyond fit indices. An in-depth description of the traditional and modern techniques for evaluating the internal structure of an assessment will be discussed. Validity evidence based on the internal structure of an assessment is necessary for building a validity argument to support the use of a test for a particular purpose. The methods described in this paper provide practitioners with a variety of tools for assessing dimensionality, measurement invariance and reliability for an educational test or other types of assessment.

Validity Evidence Based on Test Content

Stephen G. Sireci
University of Massachusetts Amherst, MA, United States
e-mail: sireci@acad.umass.edu

Validity evidence based on test content is one of the five forms of validity evidence stipulated in the Standards for Educational and Psychological Testing (AERA et al., 1999). In this presentation, we describe the logic and theory underlying such evidence and describe traditional and modern methods for gathering and analyzing content validity data. For educational tests, validity evidence based on test content is necessary for building a validity argument to support the use of a test for a particular purpose. Fortunately, many methods exist for determining how well the content of an assessment is congruent with and appropriate for the specific testing purposes.

Developing Sources of Validation Evidence across Assessment Settings

Wayne Camara
ACT, West Windsor, NJ, United States
e-mail: wjcamara@verizon.net

This presentation will review the types of evidence supporting a validation argument for assessments used in certification/licensure, clinical, educational, and employment. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. Earlier editions of the Standards for Educational and Psychological Testing focused on content, construct and criterion-related sources of validation evidence. The 1999 Standards presented a unified theory of validation and included consequential and response processes. The 2014 Standards emphasized the conceptual framework required to support each proposed use and interpretation. "Decisions about what

types of evidence are important for the validation argument in each instance can be clarified by developing a set of propositions or claims that support the proposed interpretation for the particular purpose of testing" (AERA, APA, and NCME, In press). For instance, when a mathematics college placement test is used the following propositions might be relevant: (a) which skills are prerequisite for the course(s); (b) coverage of the course content domain on the assessment; (c) test scores are supported across mathematics content represented in test items; (d) test scores are not unduly influenced by ancillary skills (e.g., writing, type of calculator); and (e) scores are related to course success across students and faculty. Examples of propositions in other testing contexts might include the proposition that test takers with high levels of consciousness exhibit such behaviors at a variety of work environments, that a certain pattern of scores associated with a clinical diagnosis are supported by evidence of similar profiles with individuals who have independently diagnosed in the same category, and that KSAs represented on a certification test are shown to be relevant and important for individuals entering that profession across environments and organizations.

Response Styles in Personality Assessment: Recent Advances

Fons van de Vijver
Tilburg University, Tilburg, Netherlands
e-mail: Fons.vandeVijver@uvt.nl

There is a new interest in the study of response styles, such as acquiescent, midpoint, and extremity responding. This invited symposium will give an overview of recent advances in this area. The contributions focus on two different approaches. The first is the implementation of novel procedures to reduce or eliminate the impact of response styles on test and survey behavior. Examples of such procedures are score standardization, the use of ipsative instruments, item calibration, overclaiming and anchoring vignettes. The second approach sets out to analyze response styles in a cross-cultural framework so as to deepen our insight in the meaning of response styles. Both types of approaches can be used to revitalize the stagnant literature on response styles and yield novel ways to deal with the old problem of response styles.

A Mixed Method Approach to the Evaluation of the Equivalence: Searching the Way for Preventing Bias in Cross-Cultural Studies

Isabel Benitez Baena[1], Fons van de Vijver[2] and José-Luis Padilla Garcia[1]
[1]University of Granada, Granada, Spain; [2]Tilburg University, the Netherlands
e-mail: ibenitez@ugr.es

The first step for comparing people for different linguistic and cultural groups in international studies is to assure bias is not present. That means that constructs, instruments and items are interpreted the same way across groups. Bias identification is especially relevant in a cross-cultural context where reaching equivalence is a requirement for making valid comparisons across groups involved. The aim of the study is to outline a Mixed Method approach for investigating the presence of bias, its causes and how to prevent it. An illustration of the proposal is shown by analysing common instruments and items for assessing well-being in international studies. In the last years, "well-being" has become one of the topics of interest due to its impact in the population health conditions, and its evaluation is habitual in international surveys. Two main phases were conducted in the study. Firstly, a quantitative evaluation of DIF was implemented by following different statistical techniques. Secondly, cognitive interviews were performed aimed to elaborate on and interpret the quantitative results. Data of adults from Spain

and from the Netherlands were analysed and collected in both phases. Quantitative analyses allowed the classification of items on base to the level of bias identified, whereas qualitative data gave information about interpretations made by participants from the two different country groups. The results integration facilitated the connection between specific elements in the evaluation and the presence of bias. Benefits of the Mixed Method approach will be discussed, as well as the utility of cognitive interviews in bias research. Future studies focused on preventing bias will be also pointed out.

Response Styles and Personality Traits: A Multilevel Analysis

Jia He[1], Dave Bartram[2], Ilke Inceoglu[3] and Fons van de Vijver[1]

[1]Tilburg University, Tilburg, Netherlands; [2]SHL Group Ltd, University of Pretoria, Thames Ditton Surrey, United Kingdom; [3]Surrey Business School, Surrey, United Kingdom

e-mail: j.he2@tilburguniversity.edu

We investigated the associations of country-level response styles with country-level personality traits. We examined the shared and unique meaning of acquiescent, extreme, midpoint, and socially desirable responding in association with the OPQ32, a forced-choice format personality measure designed to be less affected by these response styles, compared to personality inventories with Likert scales. Country-level response style indexes were derived from six waves of the International Social Survey Programme and from a meta-analysis of a social desirability scale. Country-level Personality measures were taken from various databases. In the country-level correlational analysis the four response styles formed a general response style factor which was positively associated with (1) dominance (vs. submission) in interpersonal relationships, (2) competitiveness (vs. modest and democratic) feelings and emotions, and (3) data rational thinking (enjoying to work with numbers). In a multilevel analysis, age showed a positive and education a negative effect on the individual-level general response style. Negative effects of country-level socioeconomic development and individualism and positive effects of competitiveness and data rational thinking on the individual-level response style were found. We conclude that country-level response styles are systematically associated with country personality measured by the OPQ32, suggesting that they can be viewed as having substantive meaning (i.e., culturally influenced response amplification versus moderation).

"Faking Good" on Personality Tests: Test Takers' Cognitions and the Forced-Choice Format

Anna Brown

University of Kent, Canterbury, United Kingdom

e-mail: a.a.brown@kent.ac.uk

Test takers admit to "faking good" on personality and other self-report measures (Konig et al., 2011), and the extent of faking increases when the stakes increase (McFarland & Ryan, 2000). Forced-choice formats were introduced to prevent respondents from endorsing all items. Some believe that when facing two equally desirable items, the respondent would indicate their true preference (Jackson, Wroblewski, & Ashton, 2000). Some, however, argue that direct comparison of items facilitate acute differentiation of their desirability levels and thereby increases faking (Feldman & Corah, 1960). We believe that both arguments might be correct, and that meaningful evaluation of the effectiveness of the forced-choice formats in reducing faking is not possible without understanding how the test takers think when performing comparative judgments in high stakes. The study investigates the influence of test takers' goals on cognitions when responding to personality questionnaires in high stakes. We extend the

pioneering research of Kuncel and Tellegen (2009) on test takers' cognitions to forced-choice items. This experimental study tests the effects of one between-subjects factor, "goals" (short-term employment versus long-term career choice), and one within-subjects factor, "response format" (single-stimulus versus forced-choice) on test taker cognitions. Participants respond to personality items as if their employment prospects solely depended on it. After the assessment, we ask direct questions about the truthfulness of participants' responses, and about their cognitions in a follow-up interview. A pilot study with 12 participants indicated that short-term goals elicited more desire to impress, whereas long-term career needs elicited desire to present credible yet truthful picture of the self. Both response formats prompted judgements of items' desirability. A large-scale study with N=80 young people is underway to formally test these preliminary findings; specifically, the effects of goals and response formats, as well as their interactions, on test takers' cognitions.

Construct Equivalence of the MQ across Countries and Relationships with the OPQ32r

Ilke Inceoglu[1], Alex Livesey[2], Dave Bartram[2] and Mathijs Affourtit[3]

[1]University of Surrey, Guildford, United Kingdom; [2]CEB, Thames Ditton, United Kingdom; [3]CEB, Thames Ditton, United States

e-mail: i.inceoglu@surrey.ac.uk

Much research has focused on the equivalence of personality constructs, measured as broad (McCrae & Terracciano, 2005) and narrow constructs (Bartram, 2013), demonstrating that the Big Five Factors and its sub-facets hold across diverse cultures. However, few studies have examined the equivalence of dispositional work motivation across different cultures (Inceoglu & Bartram, 2013). Work motivation might be more influenced by norms and values than personality, so the question arises whether constructs measuring work motivation are interpreted in the same way across cultures and whether relationships with personality are similar. Two studies examine the construct equivalence of a multi-construct motivation questionnaire across 17 country and language samples by comparing relationships of motivation constructs (1) within the same instrument and (2) across instruments, using a multi-construct personality questionnaire. Study 1 (N= 67659 applicants and employees) assessed within-instrument equivalence of the Motivation Questionnaire (MQ, SHL, 1992). Covariance-matrices were examined in paired samples by comparing the original UK English version with the respective country/language sample, using structural equation modelling (SEM). In Study 2 (N=15260 applicants and employees) relationships of constructs measured by the MQ and Occupational Personality Questionnaire (OPQ32r; SHL, 2013) will be examined by comparing covariance matrices between instruments in country/language pairs. Results indicate a high level of within-instrument construct (i.e. configural) equivalence in the MQ country/language samples. Study 2 is still in progress. Individuals from a range of different countries interpret the MQ constructs similar in a work context. Results of Study 2 will reveal whether equivalence is also observed in the nomological network of constructs, which tends to receive less attention in equivalence studies although such research is important for the use of instruments in international settings.

Seamless Transition to Forced Choice: Leveraging Single-Stimulus Data

Yin Lin[1], Mathijs Affourtit[1] and Ilke Inceoglu[2]

[1]CEB, Thames Ditton, Surrey, United Kingdom; [2]Surrey Business School, University of Surrey, Guildford, United Kingdom

e-mail: yin.lin@shl.com

The single-stimulus response format in self-report assessments, where respondents rate statements on a scale, is susceptible to response distortions (Birkeland et al., 2006). Such distortions are reduced when a forced-choice format, where respondents rank each group of multiple statements, is adopted (Cheung & Chan, 2002). This distortion-resistant feature is desirable when stakes are high, for example in employment tests (Jackson et al., 2000), leading to higher operational validities (Christiansen et al., 2005). Moreover, forced-choice scores are normative when Thurstonian item response theory (IRT) is applied (Brown & Maydeu-Olivares, 2011, 2013). Forced choice is thus preferred in high-stake self-report assessments. Since existing self-report assessments often have single-stimulus data readily available, two low-stake studies were conducted to investigate how such data may be leveraged when developing forced-choice measures. Respondents in the first study (N=1506) completed a 15-construct, 150-item instrument. The same statements were rated on a four-point scale and ranked in blocks of 3. These two response formats were modelled independently, using a modified version of Samejima's (1969) graded response model and the Thurstonian IRT model respectively. The resulting calibrations, when applied to scoring forced-choice responses, produced similar results (scale correlations mean 0.96, SD 0.02; individual profile correlations mean 0.95, SD 0.04). The second study validated these findings. Sample 1 (N=691) completed a 6-construct, 186-item single-stimulus instrument. A different sample 2 (N=3137) completed a shorter 60-item forced-choice version. Sample 2 was scored using relevant IRT calibrations from the two respective samples and formats. Despite the less ideal but more realistic setting, resulting scores were reasonably similar (scale correlations mean 0.90, SD 0.02; individual profile correlations mean 0.97, SD 0.05). Low-stake single-stimulus IRT calibration provides reasonable parameters for forced-choice scoring. Moreover, this comparability opens up the possibility of borrowing single-stimulus item banking methodologies to create powerful multiple-form forced-choice assessments.

Tackling Response Styles in International Survey Data: Validity Evidence from PISA 2012 for an Alternative Scoring Approach for Likert-Type Items Based on Anchoring Vignettes

Jonas Bertling and Patrick Kyllonen

Educational Testing Service, Princeton, NJ, United States

e-mail: jbertling@ets.org

Questionnaire indices based on vague response categories are routinely used in PISA and other educational large-scale assessments to measure noncognitive student factors. Responses to the underlying items partly represent construct-irrelevant response styles, especially across countries. As a result, a robust finding across all previous PISA cycles is a "paradoxical" reversal of relationships between Likert-based indices and achievement outcomes on the individual versus aggregated level. PISA 2012 introduced several new methods to tackle this problem (see Kyllonen & Bertling, 2013, for an overview). This study presents an alternative scoring approach for Likert-type questionnaire items based on anchoring vignettes (King & Wand, 2004) and provides validity evidence for the new approach to measure noncognitive constructs in cross-national settings. Findings confirm and extend earlier findings based on PISA 2012 field trial data (Bertling & Kyllonen, 2013). In the alternative scoring model numerical values for student

responses are not assigned based on the concrete response option chosen but based on the self-report answer relative to the personal standard captured by ratings of three anchoring vignettes. Several scoring alternatives are compared empirically based on data from 67 countries. Validity data for uncorrected indices is compared with anchored indices, including criterion correlations, factor structure, and relationships with response style indicators. Relationships of anchored indices with students' plausible values are highly consistent within and across countries and stronger in line with underlying theories for than for uncorrected data. Students in Asian countries show more modest response behaviour then students in western countries, leading to attenuation of their values on uncorrected indices, and consequentially to paradoxical country-level correlations. Our findings show that country league tables based on most Likert-type indices are biased by response style and that anchoring can reduce this effect.

Overclaiming Adjustment to Measure Self-Reported Mathematics Topic Exposure in PISA 2012

Patrick Kyllonen and Jonas Bertling
Educational Testing Service, Princeton, United States
e-mail: pkyllonen@ets.org

For PISA 2012 there was an interest in measuring the extent to which students were exposed to various mathematical topics in school (e.g., fractions, exponents) as a possible explanatory variable to account for mathematical proficiency differences. This was part of a larger effort to investigate the importance of "opportunity to learn." Objective measures of topic exposure (e.g., textbook surveys) were not possible, so we relied on subjective estimates, which are subject to response style bias, such as overclaiming (claiming familiarity with unknown topics). The purpose of the study was to design and evaluate a method for measuring topic familiarity through self reports while controlling for overclaiming. The measure was designed to be used, along with other indicators, to evaluate the extent to which topic exposure, or opportunity to learn, predicted mathematics achievement. Approximately 510,000 students in 65 countries were administered several measures of exposure to PISA-style item types and concepts; along with the PISA mathematics proficiency items (and reading, science, and problem-solving) (OECD, 2013). The target topic-familiarity/overclaiming measure asked students to indicate their familiarity (on a 5 point scale, "never heard of it" to "know it well, understand the concept") with 16 topics. In addition, they were presented 3 "foils" (non-existent topics). Two scores were obtained: an average familiarity rating for the 16 concepts, and an adjusted average that subtracted out the familiarity rating for the 3 foils (IRT and signal detection versions of these scores were also evaluated). Consistent with findings from the PISA 2012 pilot (Kyllonen & Bertling, 2013), we found that the bias-adjusted topic familiarity score was a stronger predictor of achievement than the unadjusted score. We also found that the bias-adjusted score increased cross-country comparability. There was a substantial negative correlation between overclaiming and proficiency. Possible explanations will be discussed.

Interdisciplinary Perspectives on International and Cross-Cultural Assessment

Bruno Zumbo

University of British Columbia, Vancouver, Canada

e-mail: bruno.zumbo@ubc.ca

This symposium will bring together leading researchers from New Zealand, England, and Canada to engage with the audience from an interdisciplinary perspective on international and cross-cultural assessment. The papers will bring to bear anthropological, international and social development, aspects of critical policy studies and the sociology of education, psychological and philosophical analyses on significant issues in international and cross-cultural assessment. The issues include: (i) how contextual and interactive aspects of testing situations impact on their meaning and outcomes; (ii) the development and articulation of a pluralist approach to test validity within a multicultural context; and (iii) the rationales driving lower and middle income countries to join international assessments as key to understanding how data validity is assembled. The closing presentation by the session chair aims to bring together broad perspective to the issues and engage the audience and presenters in a discussion.

Common Themes, Unique Challenges, and Finding Broad Perspective: Engaging the Audience in Conversation

Bruno Zumbo

University of British Columbia, Vancouver, Canada

e-mail: bruno.zumbo@ubc.ca

The first purpose of this presentation is to describe the common themes and unique challenges among the three strands of research presented in this session with an eye toward broad perspective in assessment and testing. The second purpose is to transition from the formal meeting at which which speakers deliver short addresses to an open space in which there is an interchange of ideas through engaging the audience and speakers on common themes, challenges, and opportunities raised by the three papers.

A Psychometric Approach to Test Validity: The Development of Standardised Test Materials for Maori Medium Schools in New Zealand/Aotearoa

Gavin Brown[1], Peter Keegan[1] and John Hattie[2]

[1]The University of Auckland, Auckland, New Zealand; [2]The University of Melbourne, Melbourne Victoria, Australia

e-mail: gt.brown@auckland.ac.nz

Where societies are constituted by a single group (e.g., Iceland) there are few tensions about the sociocultural characteristics of assessments. However, most wealthy societies are multi-cultural. For example, New Zealand's dominant group is the white, English-speaking European or Pakeha (65% of the population), while minority groups include the indigenous Maori people (14% of the population), people from a variety of Pacific Island nations and states (i.e., 'Pasifika') (8.6%), and Asian groups (6.6%). Given large differences between groups, it is possible that educational assessments devised by the majority group may not lead to valid interpretations about children of different groups. It is crucial that sociocultural issues related to assessments are addressed. In multi-cultural societies, three possible responses to cultural diversity are seen in how assessments are used; that is, assimilation, amalgamation or pluralism. New Zealand has adopted a pluralist approach for the indigenous Maori group. Informed by psychometric theory

about validity and evidence for interpretations and decision-making, we respond to minority group criticisms of assessment. We describe the approach taken in a computerised school testing system developed in New Zealand for students in Maori-medium schools. This system developed a culturally appropriate educational assessment that focused on independent, non-translated, development of tests and ancillary resources to benefit Maori-medium students and schools. We argue that the use of valid assessments to identify discrepancies forces us to closely re-examine schooling practices. High-quality assessments are not the cause or the solution of differential achievement; however, they can help identify where teachers and students need support to move towards more culturally effective teaching strategies, resources, and beliefs.

The Neglected Situation: Anthropological Perspectives on Testing Situations, Assemblage and Embodied Social Interaction

Bryan Maddox
University of East Anglia, Norwich, United Kingdom
e-mail: b.maddox@uea.ac.uk

This paper takes its cue from Goffman's (1964) essay on the 'Neglected Situation' to consider the how contextual and interactive aspects of testing situations impact on their meaning and outcomes. The paper describes an anthropologically orientated theory of testing situations. It draws on linguistic anthropology and socio-linguistic theory to examine the cultural and interactive production of testing situations – how they are framed as social occasions, their spatial, temporal and affective dimensions. The paper uses the work of Charles Goodwin (a linguistic anthropologist) to consider both how people interact and communicate during testing situation, including the role of social status, non-verbal communication – gesture and stance. It draws on Goodwin's work and on Actor Network Theory (ANT) to consider the special role of non-human actors in assessment practices, and the centrality of testing materials as a focus of testing situations. The paper uses Goffman and the ANT concept of 'assemblage' to consider how testing situations involve both human and non-human actors who are brought together around a shared focus of activity. These anthropological perspectives are illustrated in the paper with examples from the author's ethnographic field observations of adult literacy testing situations on Mongolia (Maddox 2014). The concluding part of the paper has an interdisciplinary focus, and discusses some possible implications and insights of this view of testing situations for psychometric assessment. It identifies opportunities and challenges that an anthropological theory of testing situations has for 'ecological' assessment frameworks and 'Third Generation DIF analysis' (Zumbo 2007). For example, how social interaction and apparently idiosyncratic contextual characteristics of testing situations might be observed statistically in the patterns of behaviour and response of groups and individuals.

What Do the Rationales for Joining International Assessments Tell us about the Production of Test Validity?

Camilla Addey
University of East Anglia, Norwich, United Kingdom
e-mail: milaaddey@yahoo.it

This paper sets out to discuss the rationales driving lower and middle income countries to join international assessments as key to understanding how data validity is assembled. The paper draws on recent literature in International Assessment Studies (including critical policy studies and the sociology of education) and on Actor Network Theory (ANT) to discuss research carried out on international assessments in Laos and Mongolia. Beyond the stated rhetorical of

‘informing policy with valid international indicators’ and access to resources, lower and middle income countries also join international assessment programmes to measure competitiveness and development progress; to serve governmentality (Foucault 1980) tactics (like the statistical eradication of policy problems); and as a global ritual of belonging to ‘be put on the global map’. In this paper I ask how these rationales feed into international assessment processes and influence the production of test validity as a subject of conceptual precision (sought by the testing community) and consensus. Based a historical analysis of validity, Newton and Shaw (2014) recommend the concept be substituted with the concept of quality in relation to measurement aims. However, this case study suggests international data and its processes further socio-political interests through the very fuzziness and malleability of the validity concept. In other words, validity is a ‘rationalizing’ feature assembled through pragmatic interests and power. I apply the analytical lens of ANT (Latour 1987) to a case study of the UNESCO Institute for Statistics (UIS) Literacy Assessment and Monitoring Programme (LAMP) in Lao PDR and Mongolia. I argue that the way validity is constructed, received, used and applied should be understood through a reformulated ‘global age’ concept of governmentality. This has implications for the multiple interpretations of validity projected and black boxed into international assessments and the conceptual and methodological foundations of such programmes.

ITC Special Sessions

International Journal of Testing: Past, Present and Future: Data Mining, and the International Aspect (Pannel Session)

Avi Allalouf[1], Frederick Leong[2], Steve Sireci[3], Stephen Stark[4], Neal Schmitt[5] and April Zenisky[6]

[1]NITE, Jerusalem, Israel; [2]Michigan State University, East Lansing, MI, United States; [3]UMASS, United States; [4]University of South Florida, United States; [5]Michigan State University, United States; [6]University of Massachusetts, United States

e-mail: avi@nite.org.il

The International Journal of Testing (IJT) is the foremost publication of the International Test Commission (ITC) and is in its 14th year of publication. The session will be devoted to several aspects of the journal's editing activity in light of past and futuretrends. It will be divided into two parts: (1) Short presentations by Pannelists, and (2) Discussion. 1. Short Presentations by Pannelists Stephen Stark (IJT Co-Editor; University of South Florida, U.S.A.), April Zenisky, IJT Associate Editor; University of Massachusetts Amherst, U.S.A.) DataMining IJT's Editorial Decisions: A Synthesis of Recent Submissions and Publications In recent years, approximately 80 to 100 unique manuscripts have been submitted annually to IJT for publishing consideration. This presentation will use submission and final manuscript decision data to provide insight into IJT's publication processes and editorial perspectives. The intent here is to describe the publication processes of IJT in the context of the full picture of what is submitted, with special attention given to illustrating content and methodological trends in publication decisions. Stephen G. Sireci(Previous IJT Co-Editor;University of Massachusetts Amherst, U.S.A.) How to be a Good Reviewer: Suggested Guidelines for Reviewers of IJT Reviewing manuscripts for scholarly journals is an important component of the educational process. The quality of the reviews not only determines whether a manuscript ultimately gets published and educates readers, but also contributes to theprofessional development of theauthors. Reviewers should provide constructive feedback to both authors and editors. Ways to provide a review that will be helpful on both fronts will be discussed. Good reviews are tactful, and in addition to improving the quality of articles, also improve the research skills of the authors. Examples of helpful and hurtful reviews will be provided with the ultimate goal of improving the review process. Neal Schmitt(Michigan State University, U.S.A.) Publication: Do IJT Decisions Differ from those of other journals? A policy-capturing study conducted by the Journal of Applied Psychology (JAP), showed that, overall, a decision to publish an article was most closely related to research design issues, data analysis concerns, the theoretical basis of the research, and the operationalization of key constructs. Among other things, this presentation suggests that it is likely that these are representative of the concerns that drive editorial decisions in general. A key element of a reviewer's decision, at least implicitly,should be based on whether a submitted paper addresses a theoretical or practical question of importance. It seems likely that the IJTpublication policy is similar to that of JAP, but with the former, there is the added imperative of weighing to what extent a paper addresses issues that are of interest to an international audience. ? Avi Allalouf (IJT Co-Editor; National Institute for Testing and Evaluation, Israel) International Journal of Testing: What Does "International" Mean? One element of the journal's mission is to promote papers that are "of interest to international audience." This raises two questions which the current and previous editors of IJT have considered in depth, and for which there is no definitive answer:(1) Who are the readers and potential readers of IJT? and (2) which topicsare of interest to "international audiences"? Some context for the editorial direction followed by IJT in the areas of content and international representation will be provided. 2. Discussion Moderator: Fred Leong (Chair of the ITC Publications Committee; Michigan State University, U.S.A.) After the four presentations,the floor will be open for comments and questions from the audience.

Advances in Computer and Internet Testing: Implications for Revising the ITC Guidelines (Pannel Session)

Dave Bartram[1], Iain Coyne[2], Ben Hawkes[3], Annalisa Rolandi[4], Anders Sjöberg[5] and Nancy Tippins[6]

[1]CEB's SHL Talent Measurement Solutions, Thames Ditton, United Kingdom; [2]University of Nottingham, Nottingham, United Kingdom; [3]IBM, London, United Kingdom; [4]Utilia HR, Verona, Italy; [5]Stocholm University, Danderyd, Sweden; [6]CEB, Greenville, SC, United States

e-mail: dave.bartram@shl.com

In July 2005, the ITC launched their International Guidelines for Computer and Internet-based testing. Aimed at test publishers, developers and users, the guidelines have become internationally recognised in highlighting good practice issues in computer-based and Internet-delivered testing and have raised awareness among all stakeholders in the testing process of what constitutes good practice. Although the Guidelines have been well-received and are making an impact both in research and practice, there is recognition that a rapid developing area such as Internet testing requires the need for regular updating of the Guidelines. For example, advances in the use of mobile devices, video game techniques, avatars and online monitoring or proctoring are not fully reflected within the current Guidelines. In addition, since they were published in 2005 we have seen the publication of the ISO Standard (ISO 10667) on Assessment Service Delivery: "Procedures and Methods to Assess People in Work and Organizational settings". This provides a potential overarching framework within which to locate guidelines focused on more specific assessment issues. This Pannel discussion is the starting point for the revision of the Guidelines and will consider issues the revised guidelines need to address. The Pannellists include the original guidelines' authors and others, all of whom have expertise in the science and practice of computer-based and Internet delivered testing. They will provide a brief statement of points they see as important and will debate current issues in Internet testing and those likely to emerge in the future. Ultimately, by understanding the issues which need to be incorporated into a set of revised guidelines, the ITC can ensure the guidelines continue to be an internationally recognised resource on best practice. Iain Coyne, University of Nottingham, UK Ben Hawkes, IBM-Kenexa, UK Annalisa Rolandi, UTilia, Italy Anders Sjoberg, University of Stockholm, Sweden Nancy Tippins, CEB, USA.

ITC Guidelines: Past, Present and Future (Pannel Session)

Dave Bartram[1], Dragos Iliescu[2], Avi Allalouf[3], Iain Coyne[4], David Foster[5], Ronald Hambleton[6] and Thomas Oakland[7]

[1]CEB's SHL Talent Measurement Solutions, Thames Ditton, United Kingdom; [2]SNSPA University, Bucharest, Romania; [3]NITE, Jerusalem, Israel; [4]University of Nottingham, Nottingham, United Kingdom; [5]Caveon, Mount Pleasant, Utah, United States; [6]University of Massachusetts, Amherst, MA, United States; [7]University of Florida, Gainesville, Florida, United States

e-mail: dave.bartram@shl.com

This session will begin with a review of the work of the ITC in developing guidelines for tests and testing. It will look back at how these were developed in the past and describe current procedures for development as well as describing those that are currently under development. In 2013, a step towards great standardisation was taken when a project was carried out to provide a uniform structure and format for the design of ITC Guidelines. Dave Bartram will present a review of the current status of ITC guidelines. Dragos Iliescu will present a review of the feedback from our recent member survey relating to guidelines from the ITC. Members of the Pannel will be

invited to comment. The session will provide opportunity for the audience to comment on current plans for future guidelines and make suggestions for topics. In particular we will welcome discussion on ways in which the guidelines can be made more useful for practice.

Issues that International Testing Organizations Need to Address (Pannel Session)

Fanny M. Cheung [1], Dragos Iliescu[2], Dave Bartram[3], Alberto Maydeu Olivares[4], Fons van de Vijver[5], José Muñiz[6], Ian Florence[7] and G. Harris[8]

[1]The Chinese University of Hong Kong, Hong Kong; [2]SNSPA University, Bucharest, Romania; [3]CEB's SHL Talent Measurement Solutions, Thames Ditton, United Kingdom; [4]University of Barcelona, Spain; [5]Tilburg University, Tilburg, Netherlands; [6]University of Oviedo, Oviedo, Spain; [7]European Test Publisher Group, United Kingdom; [8]Association of Test Publishers, United States

e-mail: fmcheung@cuhk.edu.hk

As testing and assessment are increasingly conducted across geographical borders, national and international organizations concerned with tests and testing have to address issues arising from the use of the Internet for testing, test quality standards, user qualification, language and culture issues, copyright issues, and the need for local research. The Pannel will comprise representatives from a number of such organizations, including international professional, academic, and trade associations. They will each highlight the issues of concern to them and discuss how they are tackling these challenges. The Pannellists will identify areas for co-operation among international testing organizations that could contribute to the standard of testing and assessment.

ITC Test Security Guidelines (Round Table)

David Foster

Caveon, Mount Pleasant, Utah, United States

e-mail: dfoster@kryteriononline.com

This year marks the publication of ITC's Guidelines of Test Security. The Guidelines provide direction on how to keep tests secure as they are being created, as they are administered, and as the results are used. The Guidelines also cover initial planning steps, including such concepts as proper budgeting, assigning roles and responsibilities, and physical security, making sure that tests will be secure and the resulting scores can be trusted. Finally, they provide help in how to detect a security breach and how to respond when it occurs. With a rising set of threats, bolder thieves and cheaters, and an environment where test scores are critical to educational success or job promotion, it is more important than ever for testing programs to arm themselves with knowledge and tools to protect their exams.

Contributions to Score Reporting Methods and Practices (Symposium)

Ronald Hambleton
University of Massachusetts, Amherst, MA, United States
e-mail: rkh@educ.umass.edu

Over the years, tremendous progress has been made in our approaches to the construction and analysis of educational and psychological tests, see, for example, the development of item response theory models and methods, generalizability theory, scaling methods, new item writing technologies, and the advancement of methods for validating test score use. It is only recently, that we have seen very much interest and research on the topic of test score reporting methods and practices. Of course, unless test scores are communicated clearly so that users can understand the scores and use them correctly, much of the value of technical advances is lost. But the situation is improving and there is a growing literature to inform about score reporting. This session has been organized to focus on three aspects of score reporting: Reporting of credentialing scores to enhance meaningfulness to examinees, score reporting of psychological test scores from the perspective of a test publisher, and the challenging problem of communicating subtest scores. The session will conclude with a discussion of the papers, and a look to important, future research.

Dealing with Test Score Reporting in Psychological Assessment

Pablo Santamaría
TEA Ediciones, Madrid, Spain
e-mail: pablo.santamaria@teaediciones.com

Traditionally, test developers have implicitly considered their work to be complete once the test has met the technical and psychometric requirements they set out to fulfil. The way in which the results of the test would be presented to its respondents or to professionals (score reporting) was seen as a secondary issue that did not merit further technical interest or study. Nevertheless, in recent years this assumption has been called more and more into question in the field of educational measurement (e.g., Goodman & Hambleton, 2004; Zenisky, Hambleton, & Sireci, 2009). As has rightly been said, the underlying psychometric techniques are of little worth if the results yielded by the instrument cannot be transmitted in clear and comprehensible fashion to the relevant audience. This line of research and debate has led to the inclusion of score reporting design among the technical requirements of test construction, giving rise to a range of guidelines and models for score reporting development (e.g., Hambleton & Zenisky, 2013; Ryan, 2006). Even so, these advances have scarcely been incorporated in the field of psychological assessment. The principal areas of progress as regards score reporting in educational measurement – such as an emphasis on the substantive interpretation of results and their link to intervention, analysis of the needs of the intended audience or the inclusion of pilot studies in the development of formats for the reporting of results – have had virtually no influence in the psychological assessment context. This presentation will include some reflections not only on the factors that might be holding back the incorporation of these advances in the field of psychological assessment, but also on the main guidelines and other aspects that could be of benefit to this field.

A Comparison of Methods for Examining the Psychometric Quality of Subscores

Jonathan Wedman

Umeå University, Umeå, Sweden

e-mail: jonathan.wedman@edusci.umu.se

The appropriateness of reporting subscores depends on whether they provide useful information. Subscores that fail to do so lack adequate psychometric quality and should not be reported. In this study seven methods for examining subscore quality are compared when applied to two college admissions tests, and analyses were carried out on the subtest and section levels. The section scores had adequate psychometric quality with all methods used, whereas the results for subtest scores varied. The authors recommend using Haberman's method and the related utility index because of their solid theoretical foundation and because of various issues with the other methods.

International Perspectives on Score Reporting in Credentialing: Best Practices for Providing Meaningful Feedback to Examinees

Chad Buckendahl[1], April L. Zenisky[2], Jill van den Heuvel[1] and Susan Davis-Becker[1]

[1]Alpine Testing Solutions, Las Vegas, NV, United States; [2]University of Massachusetts Amherst, MA, United States;

e-mail: drcbuck@gmail.com

Effectively communicating test results with stakeholders is a critical responsibility of test developers, and this perspective on reporting is reinforced in the International Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores (International Test Commission, 2012). One aspect of such communication involves providing feedback to examinees about their test performance that is both actionable and psychometrically supported. While feedback is perhaps more commonly present in score reports for educational tests, test results in credentialing occupy a gatekeeping role for examinees hoping to enter a profession, and so guidance for improvement aimed at failing candidates has great value in this context. Building on recent work by van den Heuvel, Zenisky, and Davis-Becker (2014), the present paper first draws on international perspectives to characterize current practices for providing meaningful feedback to candidates, by considering examples from multiple countries to highlight practical issues, and offers a synthesis of best practices in reporting test feedback from a range of testing settings applicable to certification and licensure assessment.

Present and Future of the ITC Guidelines on Test Adaptation (Symposium)

Jacques Gregoire

Catholic university of Louvain, Louvain-la-Neuve, Belgium

e-mail: jacques.gregoire@uclouvain.be

A large number of psychological and educational tests are translated and adapted across languages and cultures. Unfortunately, the quality of these adaptations can be rather poor, with harmful consequences for individuals evaluated these tests. In order to support the test adaptation quality, the International Test Commission developed Guidelines on adapting tests. The last version of the Guidelines was published in 2010 (http://www.intestcom.org/upload/sitefiles/40.pdf). The guidelines fall into four main categories: those concerned with the cultural context, those concerned with the technicalities of

instrument development and adaptation, those concerned with test administration, and those concerned with documentation and interpretation. In this symposium, the rational underlying the guidelines will be explained and their use in different countries will illustrated. Some limits and potential development of the guidelines will be discussed.

Test Adaptation and the Intellectual Property Rights

Jacques Gregoire
Catholic university of Louvain, Louvain-la-Neuve, Belgium
e-mail: jacques.gregoire@uclouvain.be

Intellectual property rights refer to a set of rights over creations of the human mind. They protect the interest of creators by giving them moral and economic rights over their creations. As tests are clearly creations of the human mind, they are covered by intellectual property rights. Most of the time the copyright does not refer to specific item contents, but to the original organization of the test (structure of the scales, scoring system, organization of the material…). Consequently, mimicking an existing test (i.e. keeping the structure of the original test and its scoring system, but creating new items), is a breach of the original intellectual property rights. When authorized to carry out an adaptation, the test developer should respect the original characteristics of the test (structure, material, format, scoring…), unless an agreement from the holder of the intellectual property allows modifications of these characteristics. In this presentation, we discuss the issue of conducting a test adaptation (i.e. developing a version of the original test taking into account new linguistic and cultural constraints) respecting intellectual property rights.

Issues in Test Adaptation in France, for Personality Questionnaires and Cognitive Batteries

Isabelle Gillet
Editions Hogrefe France, Paris, France
e-mail: isabelle.gillet@hogrefe.fr

Based on the International Test Commission Guidelines for Translating and Adapting Tests, this paper will explore the process of adaptation of cognitive tests and personality questionnaires. It is now well known that this process is far from a single translation from the source to the target language/culture. Nevertheless, despite the awareness of the original authors when they create their tests, in order to avoid homographs, colloquial expressions, idioms….for verbal items, typical representation of figurative items (mailbox, houses..), the adapter always face some tricky issues because items are samples of language, anchored in a culture, and then, vary in mental, social and cultural representations. Based on the work done on many tests adaptations, a model of "judgmental" method, which analyzes and structures the process, will be presented. This method is, of course, validated by statistical analysis performed on the data collected. Examples of these different steps will be given to illustrate the challenge of cultural transposition of the underlying psychological concepts.

Tests Adaptation and Guidelines in Spanish Speaking Countries

José Muñiz[1] and Paula Elosua[2]

[1]University of Oviedo, Oviedo, Spain; [2]University of Basque Country, San Sebastián, Spain

e-mail: jmuniz@uniovi.es

Adapting tests across cultures is a common practice that has increased in all evaluation areas in recent years. We live in an increasingly multicultural and multilingual world in which the tests are used to support decision-making in the educational, clinical, organizational and other areas, so the adaptation of tests becomes a necessity. In the Spanish speaking countries (Spain and Latin American) the original language of most adapted tests is English, although in the last years the development of national tests is increasing. The Spanish language itself has different national and regional variations, so frequently tests edited in Spanish need to be re-adapted to be used in different speaking Spanish countries. The first version and the recent revision of the International Test Commission Guidelines have been translated into Spanish (Muñiz, Elosua and Hambleton, 2013), and these guidelines are the preferred framework to carry out the adaptations. Some of the problems found in test adaptation are analyzed, such as cultural distance, and construct invariance, and current problems and future perspectives in tests and testing practices discussed.

French Adaptation of Items and of Scoring Criteria In Most Recent Wechsler's Scales

Wierzbicki Claudine

Pearson - France - ECPA, Montreuil, France

e-mail: cwierzbicki@ecpa.fr

In France, we have adapted Wechsler's scales for sixty years (Wechsler –Bellevue Intelligent Scale, ECPA, 1954), and our well-established methodology really sticks to 22 guidelines published by The International Test Commission published in 2010. To illustrate our procedures, the first part of this presentation is going to show a few examples of adaptations of verbal items and picture items, and, in a second part, we are going to insist on one important rubric of our Administration Manuals : the « Examples of response » , i.e. the construction of scoring criteria. To what extend can we lean on US Examples of response in the adaptation process? How do we anticipate the possible difficulties of scoring at the very moment of items creation? How do we proceed to create a hierarchy in 0-1-2 points, for new items and for former items through a new data collection ? We shall discuss these challenges by providing examples from our recent adaptation of the WPPSI-IV (2014).

Guidelines for Test Adaptation

Ronald Hambleton

University of Massachusetts, Amherst, MA, United States

e-mail: rkh@educ.umass.edu

The International Test Commission (ITC) is committed to improving testing practices around the world, and so, it would not be surprising for persons to know that developing guidelines for test adaptation was the first set of guidelines that the ITC set about to develop. This work began around 1991 and a first draft appeared in 1994. At that time, and much expanded today, there are hundreds of psychological and educational tests that undergo test adaptation for use in multiple languages. TIMSS adapts math and science tests into more than 60 languages; PISA works in over 30 languages; several prominent individually administered tests have been

adapted into over 100 languages; and there are hundreds more examples. Flawed test adaptations are common and a major threat to the validity of psychological and educational assessments. We believe the ITC guidelines for test adaptation have been influential with over 1000 citations. In this presentation, we will address some of the difficulties in developing test adaptation guidelines, gaining support for the guidelines internationally, and lessons learned to improve the guidelines for the future.

The Use and Control of Psychological Tests in South Africa (Pannel Session)

Tholene Sodi[1], Thandeka Moloi[2] and Nanette Tredoux[3]

[1]Professional Board for Psychology, Sovenga, South Africa; [2]Professional Board for Psychology (HPCSA), Durban, South Africa; [3]Professional Board for Psychology (HPCSA), Greenside, South Africa

e-mail: Tholene.Sodi@ul.ac.za

Pannel members Tholene Sodi (Convenor: Chair of the Professional Board for psychology) Nanette Tredoux (Member: Psychometrics Committee) Thandeka Moloi (Chairperson: Psychometrics Committee) Abstract In South Africa, the use of psychological tests is regulated through the Health Professions Act (Act Number 56 of 1974). Specifically, the Act restricts the use of psychological assessment measures to persons who are registered as psychology practitioners. Among others, the following are defined in the Act as "acts pertaining to the profession of psychology": • the exercising of control over prescribed psychological questionnaires or tests or prescribed techniques, apparatus or instruments for the determination of intellectual abilities, aptitude, personality make-up, personality functioning, psycho-physiological functioning or psychopathology (Section 2d); • the development of and control over the development of psychological questionnaires, tests, techniques, apparatus or instruments for the determination of intellectual abilities, aptitude, personality make-up, personality functioning, psycho-physiological functioning or psychopathology (Section 2e); To give effect to the Health Professions Act, and to provide proper guidance to the profession of psychology, the Professional Board for Psychology has, in the last few years, embarked on a process to develop regulations aimed at controlling the use of psychological tests in South Africa. In this Pannel discussion, three members of the Psychometrics Committee trace the development of the draft regulations. Tholene Sodi presents an overview of the draft regulations, including the mandate of Professional Board for Psychology and the challenges for developing these guidelines. Nanette Tredoux makes a presentation of the history of psychological testing in South Africa. The roles played by TCRSA and HSRC are outlined. Thandeka Moloi looks at the specific aspects of the draft regulations. These, among others, include the steps and processes to be followed in the development of tests and the statutory requirement imposed on the Board to publish (annually) a list of psychological tests.

Testing Challenges in Multicultural Contexts: Perspectives from Iberia Latin American Countries (Symposium)

Solange Wechsler

Ponticia Universidade Catolica de Campinas, Campinas, Brazil

e-mail: wechsler@lexxa.com.br

Test development and use in Iberia Latin countries face many challenges due to t various domestic conditions. In Iberia nations, their native born citizens as well as their many immigrants deserve quality psychological tests, either created and adapted to their needs. However, this standard has been achieved rarely. In South America, cultural characteristics of

each nation together with professional and legal restraints often preclude test adaptation. Therefore, each country faces the challenge to develop measures according to local conditions. International collaboration through research projects, visits, and consultation from scholars to assist human resource personnel to acquire and use needed skills while building on existing resources in these countries could have a material and beneficial impact on test development and use in these countries. This symposium explores these issues in reference to Portugal, Spain, Brazil and Argentina.

Psychological Assessment in Argentina. Raising Questions from a Cross Cultural Psychology Perspective

Evangelina Norma Contini
Universidad Nacional de Tucumán, S.M. de Tucumán, Argentina
e-mail: contini.norma@gmail.com

Psychological Assessment is a core task in the Psychologist's practice and it has led to the use of tests and diagnostic categories. A critical analysis of such practices in Argentina will be carried out on the basis of contributions by Cross Cultural Psychology; the culture-behaviour relation and the notions of universalism and relativism will be analysed. The methods used, the theories that back them up and the classifying systems of diagnosis will be referred to. The aim is to consider how the quality of diagnosis is affected by the use of tools which originate in a cultural context different from the ones they will be applied to and by the use of supposedly universal categories of diagnosis. Due to the constant need to assess subjects from different cultural backgrounds living in Argentina, this topic gains particular relevance: it is important not to mistake cultural differences for deviation or psychopathology in order to get a more reliable diagnosis which will enable more effective interventions.

Evaluating the Brazilian Movement for Test Development

Solange Wechsler
Ponticia Universidade Catolica de Campinas, Campinas, Brazil
e-mail: wechsler@lexxa.com.br

Test development in Brazil can be described by four critical phases, which have affected the current views and practices on psychological assessment. In the first phase, tests were high valued at the end of the XIX century and the use of translated tests was common among psychologists. The second phase, which lasted for several decades, tests were not held in high regard, in part, because they were imported and did not represent Brazilian culture. During the third phase, researchers and other scholars interested in psychological tests grouped together to create the Brazilian Institute of Psychological Assessment and established several university-based psychometric laboratories. In 2001, the Federal Council of Psychology enforced regulations that a psychological test can be used in Brazil only after providing empirical evidence of the adequacy of its norms and validity. In the fourth and current phase, increase in tests created in the country as well as the urge for research on validating and norming tests originated from other nations have added importantly to test development and use in Brazil. This paper discussed the advantages of this movement and the challenges to be met after 13-years this regulation. Brazil has risen to a position that allows it to help other Latin American countries in test development and use.

Tests and Testing in Spain: Current Situation and Future Perspectives

José Muñiz
University of Oviedo, Oviedo, Spain
e-mail: jmuniz@uniovi.es

An adequate use of assessment instruments requires that the tests have appropriate psychometric properties, and professionals who use those instruments have the necessary expertise to utilize them. In order to improve tests and testing we have to act on both, the instruments and the professionals. In this presentation we try to analyze the current situation of tests and testing in Spain, the main actions and projects aimed to improve testing practices will be presented and future perspectives discussed. Special attention will be paid to the evaluation of tests published in Spain, carried out with the Test Assessment Model developed by the European Federation of Psychologists' Associations (EFPA), adapted to the Spanish context. The model allows conducting a qualitative and quantitative evaluation of the tests.

Testing in Portugal

Leandro S. Almeida
University of Minho, Braga, Portugal
e-mail: leandro@ie.uminho.pt

Psychology as a discipline and profession began to emerge late in Portugal (1975), a time when an anti-testing movement had a strong voice in both Europe and elsewhere.. As a result, psychology programs did not provide adequate training in psychometry. Improvements began to occur after psychologists entered the job market and found they were unprepared for their work. For example, the main professional demands on school and mental health service providers requires good skills in psychological assessment. Lacking resources found commonly in developed countries, Portugal has had to rely on young researchers to assume this role by preparing tests as part of their master's and doctorate dissertations. Some focus on test development and others on test adaptations. These efforts generally focus on two domains: intelligence and career development measures. The country's growing number of school psychologists needs both types of tests. As a result, the country has better testing resources in these two domains than in tests for children's personality and adult assessment. This presentation will discuss these issues in more detail as well as prospects for the country's future and the possible impact of international standards on test development and use in Portugal.

Symposia

College Entrance Examination: Implementing, Modifying, and Evaluating Efficacy of Changes

Alvaro Arce-Ferrer
Pearson, San Antonio, United States
e-mail: alvaro.arce-ferrer@pearson.com

Countries around the world have embarked on finding ways to maximize college success and resources. With the ongoing education reforms and the increasingly number of students seeking higher education admission, national college admission testing programs are making bold changes to their processes and their assessments to realize the goals stated on their national policy. In this symposium, experts in the area of college admission testing and practitioners from testing agencies share their experiences of frameworks to evaluate revisions needed for a testing program, processes to revise college admission tests, and recommendations on ways to manage intended and unintended consequences of the changes. The symposium addresses important areas ranging from launching a national college admission programs, modifying pinpointed aspects of an existing national college admission program – psychometric properties of the tests – to evaluating the efficacy of the changes made to the programs. The session includes five presentations. The first paper analyzes processes followed in Iceland to launch a National Entrance Examination to replace free college admission to larger universities. The second paper discusses a study on The Swedish Scholastic Assessment Test (SweSAT) to improve test score validity and use of reported sub-scores in college admissions. The third paper deals with changes introduced in the PET – a test developed and administered by the National Institute for Testing and Evaluation of Israel– to add writing assessment, to shorten the multiple choice part, and report of additional general scores. The fourth paper discusses validation research on Turkey's college entrance examination. The fifth paper proposes a validity framework to study efficacy of modifications made to national entrance programs. This session provides the audience valuable exposure to some of the most current and interesting work to launch, modify, and evaluate college admission tests.

A Validation Framework to Study Efficacy of Changes to College Admission Programs

Alvaro Arce-Ferrer
Pearson, San Antonio, United States
e-mail: alvaro.arce-ferrer@pearson.com

College admission testing programs are modifying vehemently their processes and assessments to give answer to the growing pressures to demonstrate efficacy to deliver what they promise. National testing programs often need to make decisions on changes within very short windows of time, with limited evidence at hand, and with a conglomerate of decision-makers holding different perspectives and values. Without a declared argument developed with inputs from multiple stakeholders and decision-makers, the outcomes of change can be notoriously criticized and debated at infinitum. Evaluating the efficacy of national admission testing programs is better served by a validity argument that acknowledge technical considerations as well as the volatile arenas in which change takes place. Argument helps to articulate claims made on the program and evidence available in form of data, policy, and logical analyses; in such a way that decision-making is better served with scientific argumentation and analysis of evidence. This paper introduces a validation framework to study the efficacy of national admission testing programs. The framework goes above and beyond technical psychometric considerations on test validity (Kane, 2013; Sireci, 2012) and builds argument with an audience that may or may not share the

same set of perspectives and values. To accommodate this very distinct feature, the validation framework makes use of heuristics from Cronbach's generalizability theory, Stake's countenance model, and House's Connoisseur approach to identify universes of evaluation, their corresponding facets, and the elements characterizing them; and to build an upon agreement definition of merit and value to judge each element. The paper provides a step-by-step demonstration of the framework. Questions often arise when national entrance examination testing programs become modified to respond to national policy demanding change. The framework identifies what aspects of the program need revision, how such revision would take place, and how results would be evaluated.

The "New" SweSAT: Bold, Old, or Both?

Christina Wikstrom
Umea University, Umea, Sweden
e-mail: christina.wikstrom@edusci.umu.se

The Swedish Scholastic Assessment Test, SweSAT, was introduced in 1977 as a test for selection to higher education and it is now used alongside the upper-secondary school GPA for selection purposes. The test is optional (i.e., it serves as a second chance) and aims to measure knowledge and abilities considered important for academic performance. In 2011 the test underwent major revisions, and the main reason behind these has been to increase the predictive value of the test. Therefore, the revisions included creating two sections that could be weighted differently, one with four verbal subtests and one with four quantitative subtests. The separately scaled and equated section scores are now reported to the test takers along with the total score; still, only the total score is used for admissions (as with the old test). Furthermore, the test was deemed inefficient in terms of items per minute, and poor-balanced in terms of content given the large proportion of vocabulary items. To improve content validity and obtain an equal balance between the sections, new subtests (sentence completion, mathematical problem solving, and quantitative comparisons) were introduced and old subtests were revised. The correlation between SweSAT scores and USS-GPA is about the same as it was for scores from the older version of the SweSAT. Studies of predictive validity will be carried out once criterion data is available in early 2014. The reliabilities (α and split half) of the two sections are acceptable with rxx' around .90, and the two-factor model underlying the test score reporting provide an acceptable approximate fit to the data (CFI=.95; RMSEA=.07; SRMR=.05). Further, test score analyses and test taker surveys indicate problems with speededness, which create a challenge when determining appropriate accommodations for students with disabilities. Other challenges include score reporting and use, and scaling and equating.

PET Reform: Adding an Essay-Writing Section and More

Avi Allalouf and Naomi Gafni
NITE, Jerusalem, Israel
e-mail: avi@nite.org.il

PET is a multiple-choice (MC) test developed and administered by NITE and used for higher education admissions in Israel. It is comprised of three domains: Verbal Reasoning (V), Quantitative Reasoning (Q), and English as a second language (E). Although the PET has consistently manifested high levels of reliability and reasonable levels of validity, the public, the media and policy makers have voiced complaints about it, such as: that it is not relevant to the study programs to which students apply, and that it requires costly and intensive coaching. A need for change was recognized. The first change was the addition of a writing task: assessment

of writing ability would be a way to strengthen the link between the skills measured by the PET and the learning imparted by university curricula; it would also lead to greater awareness of the importance of writing. The writing task has been included in the PET since 2012. To maintain the same testing time frame, a second change was made: decreasing the number of items in each of the MC sections by excluding items with the lowest face validity and authenticity – mainly vocabulary and quantitative comparison items. The third change was reporting three general scores instead of one, assigning different weights to different domains. The purpose was to strengthen the link between the admissions score used and the field of study to which students apply. Many activities were entailed in the PET revision, such as creating writing tasks, developing scoring rubrics, training raters and implementing changes in the structure of MC sections. More than a year after, it seems that the goals have been achieved. There has been a positive response with regard to the writing section; reliability was affected only to a minimal degree; data is not yet available for predictive validity.

The Policies and Practice of Initiating a College Admissions Test

Sigurgrímur Skúlason
Educational Testing Institute, Reykjavík, Iceland
e-mail: sigsk@namsmat.is

The plans for a new college entrance exam in Iceland propose that its first official use will be in 2014. The talk will cover some of the issues involved in its initiation. Admission to the largest universities in Iceland have been open, anyone with a high school degree can register to almost any program. This has led to heavy teaching load on academic staff and high dropout rates. Plans of a college entrance exam as part of future admissions criteria for selected highly demanded division is met with mixed responses both within and outside the institutions. Among issues discussed will be the internal conflict within the largest universities due to expected changes in admission practice, changes in the perceived image of the institution, funding issues, teaching issues, effect on students and others. Similar conflicting reactions from political and bureaucratic authorities, students and the highs school community will also be discussed The results of experimental usage of a few test components in two and results from item tryouts in unselected student groups will be discussed. The experience of this first use of the new college admissions test, even though only three out of five planned test components were used, was positive. The program managed to decrease the number of students beginning in the fall while ending up with similar number of students passing the semester. Hopefully some results of the first full operational administration in June 2014 can be introduced.

University Entrance Examinations in Turkey

Giray Berberoğlu
Middle East Technical University, Ankara, Turkey
e-mail: berberoglu.giray@gmail.com

University admission is very competitive in Turkey because of the high demand for the university education but limited quotas in the higher education programs. In the year of 2013, 1.800.433 students applied for the university but only 877.784 of them were selected and placed in the higher education programs including the Open University. The group taking the tests is very heterogeneous in terms of curricula students exposed to and some family and student background characteristics as well. The selection and placement examinations consist of two stage tests with multiple choice questions. The first stage is basically used for screening purpose. The second stage is used for both selection and placement purposes. The whole system needs to

be reformed for a better university admission since neither students nor the educators in Turkey are content to the existing practices of selection and placement. Thus, in this presentation the entrance examination will be evaluated in terms of (1) its psychometric characteristics and (2) the school and student related factors and their relation to the success in the examination in terms of prediction power of the test scores. In the first step, the factorial validity of the subtests will be evaluated in terms of the constructs being assessed by the tests. Second, some variables such as, mode of instructional processes, high school CGPA of the students, family background characteristics, students' self efficacy and interest in subject matter areas will be considered for predicting the success in the entrance examinations. It is expected that these analyses will provide information about the properties of the test content in terms of constructs being assessed for the purpose of revising and improving the test plan.

The Assessment of Critical Thinking: Cross-Cultural and Validity Issues

Heather Butler

California State University Dominguez Hills, Carson, United States

e-mail: hbutler@csudh.edu

Information is being generated at such a rapid rate that some researchers have argued that human knowledge is doubling every year. Consumers of information must learn to think critically about the information they consume, as it is a vital part of being an informed citizen and there are con-sequences for accepting information that is not accurate. Many educators, agencies, employers, and organizations have questioned whether today's graduates will be prepared for the demands of the 21st century workplace. As a result of these concerns, much educational reform has em-phasized critical thinking instruction. The papers in this symposium examine the psychometric qualities of the Halpern Critical Thinking Assessment in several countries (the Netherlands, Por-tugal, and the USA), and the development of a new domain-specific critical thinking assessment (in Belgium). The Halpern Critical Thinking Assessment (HCTA) covers five broad categories of skills (verbal reasoning, argument analysis, hypothesis testing, likelihood and probability, and decision-making and problem solving) in the context of common everyday situations. It is the only critical thinking assessment that uses both open-ended and forced-choice response formats. Three papers in this symposium examined the validity of the HCTA. Two papers explored the real-world implications of critical thinking; more specifically, whether greater critical thinking ability was inversely related to the proportion of negative life events experienced by the respond-ents. The final paper in this symposium will introduce a domain-specific critical thinking assess-ment for physics students.

Predicting Real-World Outcomes: Is Critical Thinking Ability or Intelligence more Strongly (In-Versely) Related to Negative Life Events?

Heather Butler

California State University Dominguez Hills, Carson, United States

e-mail: hbutler@csudh.edu

Few assessments of critical thinking have been validated with "real-world" outcomes of critical thinking beyond measures of academic performance. Prior research (Butler, 2012) found a relationship between scores on the Halpern Critical Thinking Assessment (HCTA) and a behavioral inventory of negative life events. The behavioral inventory, known as the Real-World Outcomes (RWO) inventory, consists of negative life events that vary from mildly negative (e.g.,

paying a late fee for returning a movie rental late) to severely negative (e.g., filing bankruptcy). Community adults and college students (n = 285) in the United States completed the HCTA, a revised version of the RWO, and an intelligence test (the INSBAT). There was an inverse relationship between the proportion of negative life events experienced by the respondents and scores on the critical thinking assessment (r = -.324), and the intelligence test (r = -.267). That is, those with higher critical thinking scores and those with higher intelligence scores reported fewer negative life events. A multiple regression was conducted using both critical thinking and intelligence scores to predict negative life events. It was predicted that critical thinking ability would be a stronger predictor of negative life events than intelligence. Both predictors (critical thinking ability, intelligence) predicted the proportion of negative life events when entered in the model at the same time, although critical thinking ability was a stronger predictor than intelligence. Differences in the strength of these relationships between community adults and college students will be discussed.

Development and Validation of a Domain-Specific Critical Thinking Test

Dawit Tibebu Tiruneh, An Verburgh, Mieke De Cock and Jan Elen

KU Leuven, Leuven, Belgium

e-mail: dawittibebu.tiruneh@ppw.kuleuven.be

The need to embed critical thinking (CT) instruction across academic disciplines in higher education has become an important focus of researchers and educators. There have been numerous efforts in designing and implementing instructional interventions that could promote acquisition of both domain-specific and domain-general CT skills. Standardized instruments for measuring domain-general CT skills are available. However, the emphasis given to the development of do-main-specific CT skills has not been matched by the development of standardized domain-specific CT measures. The absence of such standardized measures has made it difficult to deter-mine effectiveness of interventions in stimulating the development of domain-specific CT skills. This study aims at developing and validating a domain-specific CT measure targeting a freshmen introductory physics course. We focused on the five categories of CT identified in the Halpern Critical Thinking Assessment (HCTA: Halpern, 2010): verbal reasoning, argument analysis, hy-pothesis testing, likelihood and uncertainty analysis, and decision making and problem solving. We initially examined the specific of each item in the HCTA, and constructed a set of domain-specific items targeting each of the identified CT categories. We attempted to mirror the constructed-response items of the HCTA. As each item was constructed, frequent discussions were held among the authors and with other two subject-matter experts. This phase led to sub-stantial revision of the items in which 17 constructed-response and 2 forced-choice items were finally kept. The validation involved cognitive interviews with a small group of physics students, and large group paper-pencil administration. Analyses of the data are in progress. Evidence of the validity and reliability of the test, and its future worth in examining effectiveness of instructional interventions will be discussed. It is hoped that the study will be useful for researchers who would be interested to develop domain-specific CT tests in their major areas.

Are College Students Critical Thinkers? Assessment of Portuguese College Students' Critical Thinking Using Halpern Critical Thinking Assessment

Amanda Franco
University of Minho (Braga / Portugal), Gualtar, Portugal
e-mail: amanda.hr.franco@gmail.com

A great number of universities identifies critical thinking as one of the very core pillars of their mission. Critical thinking skills have been associated to an active exercise of citizenship, higher quality decisions, and to a more efficient problem solving ability, increasing the possibility of success in daily life, academic performance, and in the labor market. Despite its multifaceted relevance, it seems that (teaching for) critical thinking is still missing in many higher education institutions, which is rather puzzling: critical thinking is presented as a key-goal of higher education, and yet, it is absent from the curriculum - considering that this set of skills and dispositions requires deliberate systematic teaching and regular practice. Still, does the lack of intentional teaching result inevitably in a lack of critical thinking, or do college students stand a chance to become critical thinkers still? To answer such questions, and to analyze the quality of regular college students' critical thinking, we assessed a sample of freshmen students taking a degree course or a master's degree, in courses integrated in the area of social sciences or science and technology, in a public university in Portugal, using Halpern Critical Thinking Assessment (Halpern, 2010). We present and discuss the results obtained according to students' academic year and study field, focusing on the aspects that discriminate freshmen degree students from master students of different scientific areas. Also, we make some considerations about the quality of Portuguese college students' critical thinking, and about the need to integrate critical thinking in the curriculum, in order to stimulate the ability to deal successfully with the ambiguity and complexity of both present and future life challenges. Key-words: Critical thinking; Higher education; Assessment; Halpern Critical Thinking Assessment.

The Halpern Critical Thinking Assessment: Towards a Dutch Appraisal of Critical Thinking

Hannie De Bie and Pascal Wilhelm
University of Twente, Enschede, Netherlands
e-mail: hannie112@hotmail.com

When implementing critical thinking learning in education, a valid and reliable instrument for assessing the level of critical thinking skills is needed. This study focuses primarily on the psychometric properties of the Dutch version of the Halpern Critical Thinking Assessment (HCTA). The HCTA was administered to university students in communication and psychology (N = 240). The mean score on the HCTA was 108.23 (SD = 13.91). Reliability of the HCTA appeared sufficient (α = .75; α?2 = .77) and factor analysis indicated that the use of the constructed response and forced choice format each containing the five categories of critical thinking skills is an adequate method for assessing critical thinking ability. Taken together, these results not only confirm the quality of the Dutch translation, but also the universality of the two factor model each containing the five subscales. Based on feedback of the respondents, we infer that complex and incorrect sentence structure and scientific questioning could have affected the comprehensibility of some of the items. Reliability may be further increased once these issues are resolved. Recommendations for improving the Dutch HCTA are discussed. Second, this study attempted to replicate the findings of Butler (2012), whose primary objective was to determine whether HCTA scores are related to a real-world outcomes inventory (RWO). Along with the HCTA, a Dutch version of the RWO (RWO-NL) was developed to measure negative life events. The number of negative life events was hypothesized to be inversely related to critical thinking ability. In

contrast to what was expected, the total HCTA and weighted RWO-NL scores (M = 0.14; SD = 0.08) did not show a significant relationship, r = -.12, ns. This may have been caused by the inclusion of only university students and unproctored administration, so future research should take these issues into account.

Psychometric Evaluations of the WISC-IV: Examinations of Reliability and Validity

Gary Canivez

Eastern Illinois University, Charleston, United States

e-mail: glcanivez@eiu.edu

Wechsler scales of intelligence are arguably among the most frequently used individual measure of cognitive abilities world-wide (Georgas, van de Vijver, Weiss, & Saklofske, 2003; Lichtenberger & Kaufman, 2009) with several translations, adaptations, and standardizations. This symposium presents a collection of papers detailing psychometric evaluations of the WISC-IV, French WISC-IV, and WISC-IV-UK. The first paper concerns the longitudinal stability of French WISC-IV scores among a non-clinical sample with a mean retest interval of two years. Most studies of intelligence test stability are conducted with disabled youths due to availability of data from periodic reevaluations. The second paper is concerned with extending understanding of the latent factor structure of the WISC–IV-UK using all 15 subtests. Only one published study is available on the WISC–IV-UK and showed support for a bifactor model of the 10 core subtests. The second paper reports on examination of measurement models from both the traditional Wechsler model as well as the contemporary CHC model. The third paper replicates and extends previous findings of the latent factor structure of the WISC-IV by applying Bayesian structural equation modeling to overcome unnecessary strict parameterization of fixing subtest factor loadings on alternate factors to zero in estimating model parameters. The final paper is concerned with determining the predictive and interpretive value of WISC–IV-UK factor index scores by examining the incremental validity of their prediction of academic achievement on the WIAT-II-UK. The importance of WISC–IV-UK factor index scores must be evaluated by what they provide beyond the FSIQ. This collection of papers helps shed light on the reliability and validity of the WISC-IV from a variety of perspectives and will supplement and compliment previous research. Importantly, these studies challenge many of the recommendations for interpretation promulgated by the publisher and various textbooks and interpretive guidebooks.

Incremental Validity of WISC–IV-UK Factor Index Scores in a Large Irish Sample

Gary Canivez[1], Marley Watkins[2], Trevor James[3], Rebecca Good[3] and Kate James[3]

[1]Eastern Illinois University, Charleston, United States; [2]Baylor University, Waco, United States; [3]Éirim: The National Assessment Agency Limited, Dublin, Ireland

e-mail: glcanivez@eiu.edu

Clinicians are directed to primarily interpret the four WISC–IVUK factor index scores (Wechsler, 2003) but internal structure studies of Wechsler scales have consistently shown that the FSIQ accounts for substantially greater portions of common and total variance. If primary interpretation of the four factor index scores promoted by the publisher is to be followed, the four factor index scores must demonstrate meaningful incremental predictive validity beyond the FSIQ. Thus, incremental validity of the four WISC–IVUK factor index scores was assessed in predicting performance on the WIAT–IIUK. Participants were 1,014 Irish children, ages 6–16,

who were referred for evaluation of learning difficulties. Males comprised a larger portion of the sample (n=635, 62.6%). The mean age of the sample was 10.77 (SD=2.57). Hierarchical multiple regression assessed proportions of WIAT–IIUK achievement test variance predicted by the observed WISC-IVUK FSIQ and factor index scores. The FSIQ was singularly entered into the first block and the four factor index scores were jointly entered into the second block. Results showed the WISC–IVUK FSIQ accounted for statistically significant ($p<.0001$) portions of WIAT–IIUK score variance ranging from 7.7% to 42.5% with mostly medium to large effect sizes. WISC–IVUK factor index scores combined to provide statistically significant ($p<.05$) variance increments in most WIAT–IIUK scores over and above the FSIQ; however, effect sizes were mostly trivial to small. Individually, the WISC–IVUK factor index scores provided trivial to small unique contributions to predicting WIAT–IIUK scores. Because WISC–IVUK factor index scores did not account for meaningful portions of achievement variance beyond the FSIQ and apportioned variance from bifactor confirmatory factor analysis of the WISC–IVUK (Watkins, Canivez, James, James, & Good, 2013) shows general intelligence dominance, primary interpretation should be of the WISC–IVUK FSIQ not the factor index scores.

Bayesian Structural Equation Modeling of the WISC-IV with a Large Referred US Sample

Philippe Golay[1], Thierry Lecerf[1], Marley Watkins[2] and Gary Canivez[3]

[1]University of Geneva, Geneva, Switzerland; [2]Baylor University, Waco, United States; [3]Eastern Illinois University, Charleston, United States

e-mail: philippe.golay@unige.ch

Numerous studies have supported exploratory and confirmatory bifactor structures of the WISC-IV in US, French, and Irish samples . When investigating the structure of cognitive ability measures like the WISC-IV, subtest scores theoretically associated with one latent variable could also be related to other factors. A major drawback of classical confirmatory factor analysis (CFA) is that the majority of factor loadings need to be fixed to zero to estimate the model parameters. This unnecessary strict parameterization can lead to model rejection and cause researchers to perform many exploratory modifications to achieve acceptable model fit. Bayesian structural equation modeling (BSEM) overcomes this limitation by replacing fixed-to-zero-loadings with "approximate" zeros that translates into small, but not necessary zero, cross-loadings. Because all relationships between factors and subtest scores are estimated, both the number of models tested and the risk of capitalizing on the chance are decreased. The objective of this study was to determine whether secondary interpretation of the 10 WISC-IV core subtests from a large referred US sample could be justified or whether a simple and unambiguous interpretation was more appropriate. Accordingly, the influence of each latent factor on subtest scores was estimated using BSEM. WISC-IV data obtained from 1,130 US children (ages 6-0 to 16-11) assessed for learning difficulties was subjected to BSEM. Two substantive cross-loadings were found with a higher order 4-factor model suggesting that secondary interpretation of some subtest scores could be adequate. However, a bifactor alternative (four first-order factors and one general factor) compared favorably to the higher-order model. With the bifactor model, no secondary interpretation of the subtest scores was supported. Results suggested a simple and parsimonious interpretation of WISC-IV subtest scores. Results also indicated that the four WISC-IV factor index scores did not necessarily provide additional and separate information from the FSIQ.

On the Myth and the Reality of the Long-Term Stability of French WISC-IV Scores

Sotta Kieng[1], Nicolas Favez[1], Jérôme Rossier[2] and Thierry Lecerf[1]

[1]University of Geneva, Geneva, Switzerland; [2]University of Lausanne, Lausanne, Switzerland

e-mail: sotta.kieng@unige.ch

Tests of intelligence are often used for diagnostics and intervention purposes. Beyond these goals, tests of intelligence are used to identify cognitive strengths and weaknesses. These diagnostic applications are based on the hypothesis that intelligence is an enduring trait. While several studies have investigated short-term stability of intelligence tests scores, few have assessed the long-term stability of tests scores. However, it is essential that diagnostics and intervention are based on stable intelligence tests scores. The objective of this study was to investigate the long-term stability of the French Wechsler intelligence scale for Children – Fourth Edition (WISC-IV) with non-clinical children. To achieve this goal, a test-retest procedure was used. The WISC-IV was administered twice to 174 non-clinical children aged from 8 to 12 years, with an average test-retest interval of 2.07 years (range 1.11 – 2.93). The long-term stability was analyzed according to interindividual stability (stability coefficient: correlation between test and retest scores) and intra-individual stability (individual difference in change). Individual changes in WISC-IV scores across the retest interval is presented within standard error of measurement. A two-standard error of measurement (±2-SEM) interval was used to assess intraindividual stability/changes of the scores. As expected, test-retest coefficients were acceptable for FSIQ, VCI, and GAI scores (r = .79, .80, and .79, respectively). This finding was consistent with the hypothesis that intelligence is an enduring trait. Regarding intraindividual stability, results indicated that agreement based on the ± 2-SEM criterion was good for GAI and PRI. This idiographic comparison demonstrated that GAI and PRI were stable for more than 70% of children. Results indicated that WISC-IV scores were relatively stable (FSIQ, GAI), and that intelligence is stable over time. The long-term stability of the other index was inadequate. Individual changes revealed that only GAI and PRI were sufficiently stable for use with children.

Latent Factor Structure of the WISC-IV-UK: Higher-Order and Bifactor Considerations with 15 Subtests

Marley Watkins[1], Gary Canivez[2], Trevor James[3], Kate James[3] and Rebecca Good[3]

[1]Baylor University, Waco, United States; [2]Eastern Illinois University, Charleston, United States; [3]Éirim: The National Assessment Agency, Ltd., Dublin, Ireland

e-mail: marley.watkins@gmail.com

Watkins, Canivez, James, James, & Good (2013) found support for a bifactor measurement model with the 10 core WISC–IV(UK) subtests in a large sample (N=794) of Irish children referred for evaluation of learning difficulties. Omega hierarchical coefficients showed substantial reliability of the general intelligence dimension but inadequate reliability of the four specific dimensions (VC, PR, WM, PS). No studies have been published examining the construct validity of the WISC–IV(UK) using all 15 subtests that would allow for examination of rival CHC measurement models. This study examined the latent factor structure of the 15 WISC–IV(UK) subtests with data obtained from evaluations to assess learning difficulties of Irish children. One, two, three, four (Wechsler), and five (CHC) oblique first-order factor models were examined along with Wechsler based higher-order and bifactor models and CHC based higher-order and bifactor models through CFA. Preliminary analyses found the oblique four–factor (Wechsler) and oblique five–factor (CHC) models fit better than one, two, or three oblique factors but did not meaningfully differ from each other in fit statistics (x2, CFI, RMSEA, AIC) using contemporary standards. Meaningful differences between Wechsler and CHC based higher-order and bifactor

models were not observed. For statistical and theoretical reasons, the bifactor models provided the best explanation of the 15 WISC–IV(UK) subtest factor structure as found in other investigations (Canivez, 2013; Gignac, 2005, 2006; Watkins, 2010; Watkins et al., 2013). Like previous studies (Canivez, 2013; Watkins et al., 2013), omega hierarchical coefficients from the Wechsler and CHC based bifactor models were substantial for the general dimension in both the Wechsler and CHC but inadequate for the 4 (Wechsler: VC, PR, WM, PS) or 5 (CHC: Gc, Gv, Gf, Gsm, Gs) specific factors. Implications for interpretation of WISC–IVUK scores are discussed.

Tests of Literacy and Numeracy in Multi Lingual Education: Issues in the Philippines

Esther Care

University of Melbourne, Parkville, Australia

e-mail: e.care@unimelb.edu.au

In this symposium, the development and implementation of tests administered to 900 students aged 4-6 years old on a one to one basis in the Autonomous Region of Muslim Mindanao in the Philippines, is discussed. The research program was developed to investigate the learning trajectories of students attending very different learning environments, and to identify the comparability of educational outcomes across these environments. Half of the students were attending learning centres operated by the non-government organisation, BRAC, and half were attending state run elementary schools. The first paper (Vista) provides an outline of the context in which new tests of learning outcomes are being developed to monitor and support student progress. The characteristics of the numeracy and literacy curricula that identify student skills which are to be built over a period of years were identified and provided the blueprint for the test development. In the second paper, Preclaro presents the process and its rationale for the translation of tests from the base English version first developed using the curriculum (written in English). Finally, Care discusses the characteristics of test items in the context of student groups attending different learning contexts, across different provinces in which the language of instruction varies, and across two year levels from Kindergarten to Grade 1. The study identifies the myriad challenges in implementing and assessing a mother tongue based multi lingual education in a system where currently 19 languages have been approved for education delivery through the state-based education system. In particular, the study raises questions concerning the transfer of English language based assessment approaches in early literacy.

A Study of Different Performance on Achievement Tests Across School Context and Language/Province

Esther Care

University of Melbourne, Parkville, Australia

e-mail: e.care@unimelb.edu.au

The literacy and numeracy tests were piloted across three languages - Maranoa in Lanao del Sur, Maguindanaon in Maguidanao, and Sama in Tawi-Tawi with approximately 480 students providing data from each of the three provinces. The pilot assessments were undertaken: to check their viability in terms of logistics such as test time and administration skills required; to analyse how the test items behaved across language and grade level; and to determine the dimensionality of the tests. The Kinder and Grade 1 tests were piloted in September 2013 prior to their intended use in a longitudinal research and evaluation project the following year. Items

were entered into Conquest (Adams & Wu, xxxx) to identify which did not exhibit good fit to the Rasch model, which displayed differential functioning (DIF), and whether the difficulty levels of the items correspond with a developmental learning paradigm which the curriculum reflects. The major factor of interest was province since one of the likely causes of difference would be language. The item fit was excellent with overall reliability of .95 for both numeracy and literacy. Some differences across provinces in difficulty level on particular topics in numeracy, such as time, was demonstrated. In literacy there was variation across constrained skills such as letter knowledge, reflecting the different lexicons. Some 10% of items did not display good fit, attributable to poor design, poorly understood stop rules by test administrators, and some issues of dependence. Notwithstanding major lexicon differences across the languages into which the tests were translated, the tests performed similarly across provinces. This outcome can be attributed to the method of translation used, the emphasis on item development congruent with curriculum objectives, and training of test administrators.

Development of Achievement Tests to Identify Student Learning Outcomes in Alignment With the Philippines Education Reform K – 12

Alvin Vista

University of Melbourne, Parkville, Australia

e-mail: vistaa@unimelb.edu.au

This paper presents a blueprint for developing a set of assessment tools to monitor and support student progress across the primary years of basic education. The objective is to develop a set of achievement tests that measures numeracy and literacy outcomes for both public school students and out-of-school children and youth serviced by an alternative delivery education system. For these measures to be interpretable within the Philippine educational framework and be set on a formal progression scale that follows the span of the educational system, the tests need to be aligned with the national curriculum. The design of the tests was grounded by first, their alignment with the national curriculum and its coverage from broad domains to more specific content standards and learning competencies defined by the Philippine Department of Education. Second, the tests must satisfy logistical requirements unique to the target population. This entails being amenable to multiple translations and user-friendly to the test taker and test administrator rural and low SES settings. The method 1) Specifies the components of the Philippine curriculum for numeracy and literacy; 2) Develops and adapts existing items mapped to the curriculum; 3) Reviews the item pool in terms of curriculum coverage, contextual appropriateness, and logistical considerations; 4) Designs a uniform layout and structure for the set. The numeracy and literacy tests were developed, produced, and administered to the target populations in Mindanao, Philippines. These Kinder-Grade 1 achievement tests for the 4-6 age group represent the first level in what will eventually cover the primary years of basic education. The development process provided important lessons in developing assessment tools for diverse student populations. In particular, balancing psychometric needs, educational policy requirements, and practical limitations revealed the complexity of test development in non-English speaking background and developing countries.

Translation Processes in the Development of Achievement Tests Across Three Languages – Maranoa, Maguidanaon, and Sama

Maria Hazelle Preclaro
University of the Philippines, Quezon City, Philippines
e-mail: hz.preclaro@gmail.com

The languages of instruction in Philippines in recent times have been English and Filipino. In accord with the K-12 reform, the language of instruction can now be any of 19 Department of Education approved languages. This requires that assessment tools reflect the language of instruction. The tests developed for this project in alignment with the curriculum needed to be translated into the mother tongues used as the languages of instruction in the targeted schools. Given lack of lexicons for all languages this poses some challenges. The method adopted for translation was of multiple forward translations, rather than a traditional forward, then backward approach. In the first step, for each language two linguists translated the tests in consultation with the test development teams for numeracy and literacy. The duplicate translated tests for each language were then taken to educators who were native speakers of the languages to "harmonise", taking into consideration daily language usage. The process of translation was particularly difficult for the literacy test. The characteristics of the English language that provide ease of access to early skill acquisition in English are not similarly reflected in languages such as Maguidanaon, Maranao, and Sama, which are noticeably more polysyllabic than English. Test administrators of the pilot version of the tests were provided with instructions which facilitated collection of qualitative data to inform and supplement the quantitative data analyses. The approach of multiple forward translations harmonised by daily users of the language appears to be promising. The translations were focused on equivalence of intent rather than strict translated equivalence. Given the particular characteristics of the destination locations for assessment, this approach bore fruit in terms of reflecting known objects and activities to act as stimulus objects for the students to demonstrate their literacy and numeracy skills.

Personality Assessment with the Combined Etic-Emic Approach: Recent Applications of the CPAI-2 & CPAI-A and their Incremental Validities

Weiqiao Fan
Department of Psychology, Shanghai Normal University, Shanghai, China
e-mail: fanweiqiao@shnu.edu.cn

Both cultural universality and cultural relevance are important considerations for the validity of personality assessment. Particularly, personality assessment in non-Western cultures has traditionally relied on the importation of western personality tests and neglected the cross-cultural relevance of imposed-etic measures. However, overemphasis on cultural uniqueness suffers from the marginalization of the indigenous approach, limiting cross-cultural comparisons. The value of indigenous measures should be assessed in terms of their incremental validity. The Cross - cultural (Chinese) Personality Assessment Inventory (CPAI) was developed using a combined etic-emic approach to cover both universal and indigenous personality constructs, originally for use in the Chinese cultural context. An indigenously-derived personality factor tapping interpersonal relatedness distinct from universal personality factors was identified in the CPAI. Based on cross-cultural applications of the CPAI, the CPAI 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 Western-based universal

personality structure. This symposium presents the latest research on the CPAI-2 for adults and CPAI-A for adolescents from China, Hong Kong, and the United States. The symposium comprises of presentations on the psychometric characteristics of the CPAI-2 or CPAI-A, and on the contributions of universal personality scales and indigenously derived personality scales in understanding career development and behavior, and future time discounting.

Interpersonal Relatedness and Future Time Discounting among the Chinese

Yiqun Gan[1], Xiaolu He[1] and Fanny M. Cheung[2]

[1]Peking University, Beijing, China; [2]Chinese University of Hong Kong, Hong Kong, Hong Kong

e-mail: ygan@pku.edu.cn

Previous research has largely focused on the influences of impulsivity and affect on temporal discounting; however, other personality traits may also interact with environmental cues to determine temporal-discounting choices. The purpose of the present two studies was to test the hypothesis that interpersonal relatedness (IR) affects temporal discounting differently in different interpersonal situations. Thus, we examined the interaction between IR and interpersonal rejection. Interpersonal rejection was primed prior to the presentation of the experimental task for one group of participants (the ostracized group); a control group received neutral priming. After priming, in Study 1, 111 participants performed a delayed-discounting task involving stimuli that varied in terms of reward value and immediacy; in Study 2, 57 participants performed a prioritizing task involving stimuli that varied in importance and urgency. The results indicated that IR interacted with priming group to predict discounting rates (Study 1) and the number of irrational choices made (Study 2). For the control group, the level of IR did not significantly affect participants' discounting rates or number of irrational choices. However, for the ostracized group, participants with high IR demonstrated significantly higher discounting rates or a greater number of irrational choices than those with low IR. These results indicate that the impact of IR on temporal discounting varies as a function of interpersonal stress.

How Personality Influences Career Related Efficacy Perception—The Contribution of Perceived Academic Performance and Parental Support

Jiaying Wang[1] and Fanny M. Cheung[2]

[1]The Chinese University of Hong Kong, Hong Kong: [2]Department of Psychology, The Chinese University of Hong Kong, Hong Kong

e-mail: nanfoud@gmail.com

The current study explored how universal and culture-specific personality dimensions influenced students' career related efficacy perception, including career decision self-efficacy (CDSE) and collective efficacy, and how those influences were mediated by intrapersonal resource of perceived academic performance and interpersonal resource of parental support. Using Cross-cultural (Chinese) Personality Assessment Inventory for Adolescents (CPAI-A; Cheung, Leung, & Cheung, 2005), data from 613 high school students of Hong Kong and 776 high school students of Zhejiang in mainland China, indicated that for both samples, universal personality dimensions (Social Potency, Dependability) were related to both career decision self-efficacy and collective efficacy, whereas culture-specific personality dimension was related to collective efficacy. Regional difference was also found that among Hong Kong students, culture-specific personality, Interpersonal Relatedness (IR), can only contribute to career decision self-efficacy indirectly via collective efficacy, but among students from Zhejiang, Interpersonal Relatedness also had a direct effect on career decision self-efficacy even after controlling the effect of collective efficacy. In both samples the influence of universal personality was only mediated by intrapersonal

resource—perceived academic performance, and the influence of culture specific personality was only mediated by interpersonal resource—parental support.

Career Decision Self-efficacy as a Predictor of Vocational Identity among three Chinese Regional High School Samples—Moderating Effects of Relational Personality and Collective Career Efficacy

Sarah Lai Yin Wan[1] and Fanny M. Cheung[2]

[1]The Hong Kong Institute of Education, Hong Kong, Hong Kong; [2]The Chinese University of Hong Kong,

e-mail: sarahwan@ied.edu.hk

This study investigated the role of career decision self-efficacy (CDSE) in predicting vocational identity (VI) among three Chinese regional samples (577, 719 and 711 high school students in Hong Kong, Shanghai and Zhejiang, respectively). Using Cross-cultural (Chinese) Personality Assessment Inventory for Adolescents (CPAI-A; Cheung, Leung, & Cheung, 2005), this study also examined the moderating effects of relational personality (i.e., Interpersonal Relatedness) and collective career efficacy on the relationship between CDSE and VI. Results from additive regression models indicate that CDSE and collective career efficacy related positively to VI in all regional groups. Significant moderating effects of collective career efficacy on the association between CDSE and VI were found in the Hong Kong and Shanghai samples. In particular, CDSE was less strongly associated with VI among Hong Kong and Shanghai students who showed higher levels of collective career efficacy. In addition, direct effects of Interpersonal Relatedness on VI and moderating effects of Interpersonal Relatedness on the association between CDSE and VI were only found in the Zhejiang sample. Specifically, CDSE was less strongly related to VI among Zhejiang students who reported higher levels of Interpersonal Relatedness. Present findings support the cultural relevance of the relational factor of CPAI-A and the collective career efficacy for studying relationships between personality, career self-efficacy and vocational identity in high school students in collectivistic cultures.

Testing the CPAI-A across Cultures: Hong Kong, Mainland China, and the USA

Xiaolu Zhou[1], Weiqiao Fan[1], Fanny M. Cheung[2] and Yaoming Gao[1]

[1]Shanghai Normal University, Shanghai, China; [2]The Chinese University of HongKong, Hong Kong,

e-mail: zhouxiaolucn@gmail.com

This study examined the factor structure of the CPAI-A, the adolescent version of the CPAI, across three cultural groups --- Chinese, Hong Kong, and US adolescents. The CPAI-A was standardized among Hong Kong adolescents in 2006 and among Mainland China adolescents in 2008. MIMIC models (Multiple Indicators/Multiple Causes) allow researchers to identify invariance by testing the effects of cultural group membership. In the present paper, 409 Chinese, 529 Hong Kong, and 212 the United States participants filled out CPAI-A. Two phases of multi-group modeling were analyzed. The first phase tested the scale's factor structure without controlling for the demographics differences between the groups. The second phase included the culture variable. Results indicate that the CPAI-A is factorially invariant, but some culturally-relevant differences were found as well. Implications on the use of MIMIC models in the personality assessment are discussed.

The Development and Validation of the Short Form of the Cross-Cultural Personality Assessment Inventory (CPAI-2)

Mingjie Zhou and Jianxin Zhang

Institute of Psychology, Chinese Academy of Sciences, Beijing, China

e-mail: zhoumj@psych.ac.cn

The Chinese Personality Assessment Inventory (CPAI) (Cheung et al., 1996) was developed as an omnibus indigenous personality inventory for the Chinese people using a combined emic-etic approach. Findings from research in North American and other Asian countries suggest that the CPAI is cross-culturally relevant, and the indigenously derived constructs are not restricted to the Chinese context. The CPAI-2 has been renamed Cross-Cultural Personality Assessment Inventory-2 to reflect the broader scope cultural relevance of the instrument. Although CPAI-2 has demonstrated good psychometric qualities and has been widely used by researchers, its large number of items restrict the extensive use of the measure. Based on this, the research team aimed to develop a short form of CPAI-2, by reducing the number of items whilst retaining high validity and reasonable psychometric qualities. Samples used for item reduction included the original normative sample in addition to a new representative sample (N= 2510). We selected two items from each scale based on item-total correlation and the rational content consideration. We changed the resulting 56 true-false items into 5-point Likert scale items. Packing the two items of every scale into one value, we examined the factor structure using EFA and CFA in two independent samples. Results showed that near identical four factor structures with CPAI-2 were demonstrated in EFA. Meanwhile, CFA confirmed that the short form obtained an acceptable goodness of fit. The short form also showed relatively satisfactory reliability for each factor, and average size correlations with some criterion variables (e.g., Five Factor Model, mental health, work engagement, job satisfaction). Therefore, the short form provides a practical and valid alternative for the original CPAI-2 when administration time is limited.

Investigations of Possible Cheating on High Stakes Tests

David Foster

Caveon, Mount Pleasant, Utah, United States

e-mail: dfoster@kryteriononline.com

Quote from ITC Guidelines "Unfortunately, cheating cannot be completely prevented, even with monitoring and other deterrents in place; the temptation to cheat can be immense, especially in high stakes testing." (ITC, 2011) The wisdom of this observation by the ITC in 2011 has been repeatedly demonstrated in testing programs throughout the world. Reference INTERNATIONAL TEST COMMISSION (2011). ITC GUIDELINES FOR QUALITY CONTROL IN SCORING, TEST ANALYSIS, AND REPORTING OF TEST SCORES. [HTTP://WWW.INTEST.ORG]. This symposium will help attendees evaluate the extent to which their testing programs are properly prepared to determine the need for test security investigations, to carry out professionally sound investigations when they are warranted, and to make appropriate use of the results. The following issues will be reviewed: • Developing Sound Test Security – Development process and dissemination plan • Reviewing and Possibly Upgrading Your Test Security Agreement – A critical element of planning a program and managing an investigation • Statistical methods of detecting plagiarism and cheating • Determining Whether and What Type of Investigation is warranted • Defining Roles and Responsibilities for Test Security Investigations • Adhering to all Laws and Regulations • Collecting and Evaluating Evidence • Arriving at a Decision and Managing an

Appeal Process Attendees at this symposium with be given an opportunity to share their own experiences with test security investigations and the Pannel will help the group draw lessons from them. Symposium Participants • Ardeshir Geranpayeh, Ph.D., Head of Psychometrics for one of the largest testing programs in the world • Marc J. Weinstein, Attorney at Law, experienced at the legal and investigative phases of evaluating test cheating cases and initiating legal actions • David Foster, Ph.D., Co-editor of Handbook of Test Security, eleven years full time experience helping testing programs prevent and detect cheating as well as carry out follow up investigations.

Investigations of Possible Cheating on High Stakes Tests

Ardeshir Geranpayeh

University of Cambridge, Cambridge English Language Assessment, Cambridge, United Kingdom

e-mail: geranpayeh.a@cambridgeenglish.org

Cheating has become a very serious problem in schools and colleges alike. Cheating can be seen as any action that violates the rules for administering a test. Cheating can take a wide variety of forms and may involve any of the stakeholders in the testing process including candidates, their teachers and those responsible for administering the test. Cheats may be motivated by material rewards such as access to life chances or by personal needs such as competitiveness or a lack of self-confidence. Whatever the cause, teachers, examination boards and their agents clearly have a responsibility to discourage cheating on their tests and to minimise it wherever possible. In this paper, we argue that cheating will directly question the validity of a test. Any single examination score obtained by fraudulent means is not valid; it cannot be interpreted as a fair reflection of the candidate's abilities. Results obtained through cheating have a negative impact on the validity of scores obtained by other candidates. When access to university places or employment opportunities is limited, the candidate who succeeds through fraud denies these opportunities to others. Where cheating is seen to be widespread, even honestly obtained test results may lose credibility and certificates become devalued. We review the cheating concept, its manifestations, and consequences more broadly, and focus extensively on various psychometric techniques for the detection of various forms of cheating in high stakes standardized language assessments. The paper concludes with the review of the standards for the prevention of cheating and report on various measures to control this phenomenon and report on the policies on punishment as a deterrent mechanism to minimise its impact.

Investigative Methods in Response to Possible Cheating on High Stakes Tests

Marc Weinstein

Dilworth Paxson LLP, Philadelphia, Pennsylvania, United States

e-mail: mweinstein@dilworthlaw.com

With near daily reporting in the media about cheating scandals in connection with high stakes assessment, licensure and certification tests administered throughout the world, interest in test fraud and cheating is at an all-time high. Now, more than ever, test publishers must implement and enforce vigilant test security policies and take decisive action against any person who seeks to undermine the integrity of their tests. However, test publishers can only achieve these goals by ensuring that they have comprehensive test integrity programs and that they conduct thorough and effective test security investigations that enable the test publisher to pursue all legal remedies available to them against the perpetrators of fraud. This presentation will highlight case studies and best practices for investigating potential test fraud and explain how to better recognize preliminary signals of potential test fraud. The presentation will further describe

effective investigative methods to follow up on known signals of test fraud to determine whether further investigation is warranted, identify the most important evidence to gather and describe recognized methods of gathering and preserving the evidence. Finally, this presentation will highlight various legal remedies available to test sponsors against perpetrators of test fraud or theft of test items, following the completion of test security investigations.

Assessing your Readiness to Conduct Test Security Investigations

David Foster

Caveon, Mount Pleasant, Utah, United States

e-mail: dfoster@kryteriononline.com

This presentation will focus on helping attendees evaluate the extent to which their testing programs are properly prepared to determine the need for test security investigations and to carry out such an investigation. Best practices will be identified in each of the following areas: • Developing Sound Test Security – Development process and dissemination plan • Obtaining Management Support for Conducting Investigations and Acting on the Results – What actions will you be able to take if misbehavior is confirmed? • Defining Roles and Responsibilities for Test Security Investigations • Providing Training for Conducting Investigations • Collecting and Evaluating Evidence • Arriving at a Decision and Managing an Appeal Process In each instance, examples will be provided of productive and not-so-productive ways that testing programs have dealt with these issues. Attendees at this symposium with be given an opportunity to share their own experiences with test security investigations and the Pannel will help the group draw lessons from them. Attendees should leave this symposium with a perspective on how ready their testing program are is to plan and carry out test security investigations and to act productively on their results.

Assessment and Teaching of Twenty First Century Skills

Patrick Griffin

University of Melbourne, Parkville, Australia

e-mail: p.griffin@unimelb.edu.au

Assessing outcomes associated with twenty first century skills posed challenges. Many skills are complex and currently ill-defined (such as collaborative problem solving). Assessment of these learning outcomes may require major revisions and refinements of the methods od development. This symposium examines these issues, outlines some of these 21st century skills and outlines how measurement theory is currently being applied to the design of assessment instruments to assist with the design of appropriate learning activities. Patrick Griffin will provide an overview of the project in paper one. In paper two, Esther Care and Claire Scoular provide a description of the development of collaborative problem-solving tasks using human to human interaction via the Internet. In paper three Nafisa Awwal and Claire Scoular describe the systems architecture required for human to human interaction and the way in which responses or monitoring logs files are used to identify indicators of social and cognitive behaviour among the students participating in the collaborative problem-solving tasks. In paper four Susan Harding and Patrick Griffin describe the calibration of bundles of tasks that yield estimates of social and cognitive skill levels of students. They also describe how described scales or developmental continua of social and cognitive development were defined by the data. All of this was yielded by an automatic coding and scoring procedure prior to the calibration process. The symposium

presents the conceptualisation and outcomes of a five year project. It illustrates how curriculum content-driven design specifications can be supplemented by scientific reasoning development amongst students. It also illustrates that human to human interaction by the internet is possible for the development of social and cognitive skills among students and how the definition of those developmental progressions can help to inform teachers to improve teaching and learning of 21st century skills.

The ATC21S Project

Patrick Griffin

University of Melbourne, Parkville, Australia

e-mail: p.griffin@unimelb.edu.au

This paper describes the overview of a five year project supported by Cisco Intel and Microsoft to develop 21st century skills that could be both assessed and enabled teaching or instruction in specific skills pertinent to the 21st century. The project set out to define, design and disseminate materials for the assessment and teaching of C21 skills. The ATC21S project had a research and development plan that consisted of five phases: conceptualisation, hypothesis formulation, development, calibration and dissemination. Within the conceptualisation ¬phase, the project focused on the definition of 21st century skills. The paper outlines the selection and conceptualisation of the skills to be assessed. It describes how this led to the development of hypothesised learning progressions which portrayed how the skills might vary across more and less adept individuals. Assessment tasks were commissioned to be developed from a mixture of commercial agencies and universities. The collaborative tasks were then subjected to concept checking, cognitive laboratories, pilot studies and calibration trials. The final stage of the process was dissemination, which includes the development of scoring, reporting and teaching in the Assessment and Teaching of 21st Century Skills system. Assessing collaborative behaviours and problem solving skills was possible suing human to human interaction over the internet.

Specifications for Collaborative Problem Solving Assessment

Esther Care

University of Melbourne, Parkville, Australia

e-mail: e.care@unimelb.edu.au

This paper outlines two distinct types of collaborative problem solving tasks; curriculum independent and curriculum dependent, each allowing students to apply different strategies to solve problems. The project identified procedures, specifications and templates that enable human to human interaction via the Internet in collaborative problem solving tasks based on either curriculum or hypothetico deductive reasoning. Curriculum independent tasks were developed to emphasise the enhancement of inductive and deductive thinking skills. Curriculum-dependent tasks allowed students to draw on knowledge gained through traditional learning areas or subjects within the curriculum. The collaborative problem solving construct emphasised communication for the purpose of information gathering, identification of available and required information, identification and analysis of patterns in the data, formulation of contingencies or rules, and generalisation of the rules and test hypotheses through challenging these. Characteristics of tasks which were identified as appropriate for eliciting collaborative problem solving processes are reported and illustrated by exemplar items. Collaborative problem solving tasks can be designed involving human to human interaction and requiring either hypothetico-deductive reasoning or curriculum content. The advantages and disadvantages of these two types of items are discussed.

Calibration of Collaborative Problem-Solving Task Bundles

Susan-Marie Harding
The University of Melbourne, Parkville, Victoria, Australia
e-mail: s.harding@unimelb.edu.au

This paper describes the calibration and interpretation of indicators that were identified as being sufficiently stable to be used for defining the underpinning construct. The analysis identified a set of indicators using one parameter and two parameter item response models. The study also explored the need for sufficient data points for interpretation and explored the stability of the task indicators across countries. The initial analysis identified sets or bundles of collaborative problem solving tasks that provided sufficient information to estimate student ability and provide an interpretation of the student's readiness to learn. Item parameters were then explored across countries to examine cross country stability and to support the hypothesis that the tasks were measuring the same skills in each of the participating countries. Item difficulties and thresholds were identified within countries and a series of scatter plots were used to examine the stability of item parameters across countries. One and two parameter models were deployed to identify the item estimates allowing slopes to vary in one analysis and fixing the slopes in another. The fixed slopes were set in order to interpret the underlying construct. Allowing the slopes to vary obtained a better fit of the data to the model. Approximately 150 indicators were identified that could be described as being sufficiently stable data to fit the one and two parameter models. Scatter plots also indicated that there were very few outliers of item difficulties across countries. The analysis indicated that collaborative problem-solving skills were stable across the six countries and provided sufficient data to interpret student development in social and cognitive dimensions.

System Architecture and Coding for Collaborative Online Assessment

Nafisa Awwal and Claire Scoular
The University of Melbourne, Carlton, Australia
e-mail: n.awwal@unimelb.edu.au

This paper examines the system architecture required for human to human interaction in collaborative problem solving. It establishes a procedure for defining a scoring process which enables calibration and scoring together with links to the reporting of individual student results for teachers to use in the classroom. This aspect of the study involved the identification of indicators that were interpretable and allowed automatic scoring of the collaborative problem solving tasks. The procedure began with the identification of task features that matched elements of the skills frameworks, followed by the generation of simple rules to collect data points to represent these elements. The data points were extracted from log files generated by students engaged in the assessment tasks. The data points documented each event, chat, and response made by each student. The process of coding, scoring, calibration, interpretation and reporting links is described. The paper includes examples of the process for defining and generating global and local (task specific) indicators, and examples of how the indicators are coded and scored. Approximately150 indicators were identified and these were used to map the student performance on to the framework outlined in paper two of this symposium. These 150 indicators were then used to identify social and cognitive dimensions of collaborative problem solving at an individual student level. It was possible to use human to human interaction via the Internet in collaborative problem solving to identify social and cognitive dimensions of problem solving in a collaborative manner. The implications of the project and its developments are discussed.

Exploiting Potentials of Behavioral Process Data from Computer-Based Testing

Frank Goldhammer

DIPF - German Institute for International Educational Research, ZIB - Centre for International Student Assessment, Frankfurt am Main, Germany

e-mail: goldhammer@dipf.de

Computer-based testing provides new insights into behavioral processes of item and test completion that cannot be uncovered by traditional paper-based instruments. A testing system can automatically record log file data including detailed information on what happened during the process of interacting with a number of different stimuli and questions. This enables to assess cognitive abilities and underlying processes not only by means of the outcome of an item but also by means of behavioral process data. This symposium aims at highlighting the potentials of process data by connecting it to various substantive and methodological research questions. The first contribution by Kretzschmar, Müller, and Greiff investigates for the domain of complex problem solving the heterogeneity of exploration behavior and evaluates empirically a new conception of successful exploration strategies. The second contribution by Kröhne, Goldhammer, and Hahnel addresses the comparability of paper-based and computer-based assessment of reading with respect to navigation process and timing behavior. The third contribution by Reips focuses on the analysis of behavioral process data from Internet-based testing and will address the validity of several behavioral process collection techniques. The final contribution by Goldhammer, Kröhne, and Hahnel considers individual differences in behavioral processes as a threat of construct validity and investigates how the timed administration of basic reading skill items affects validity. This symposium highlights the potential of exploiting behavioral process data for various assessment domains and purposes. This includes conceptualizing process-related constructs, building process indicators from log file data, relating individual differences in processes to outcomes, and finally, shaping the process of item completion by means of computer-based test administration. To this end, it connects to a research field of high practical relevance with a number of implications that is in need of rigorous empirical research as presented in this symposium.

Timed Administration of Items Increases Convergent Validity: Examples from Word Recognition and Sentence Verification

Frank Goldhammer, Ulf Kröhne and Carolin Hahnel

DIPF - German Institute for International Educational Research, ZIB - Centre for International Student Assessment, Frankfurt am Main, Germany

e-mail: goldhammer@dipf.de

A major characteristic of the test taking process is how much time individuals spent on items. The speed-ability tradeoff suggests that a test taker may increase speed at the expense of exhibited ability and vice versa. This becomes a measurement problem if there is between-person variation in the speed-ability compromise. Then the ability measure is confounded with the decision on speed which questions the validity of the measure. The major goal of this study was to compare convergent validity of untimed and timed measures of word recognition and sentence verification with respect to reading competence. A total of N = 888 students of 15 years old participated in the German computer based add-on study of PISA 2012. For timed administration, the response-signal (RS) paradigm was used requiring an immediate response as soon as an acoustic signal was presented. Tasks were completed both in an untimed condition allowing individual differences in time spent on task, and in a timed condition where the time available for item completion was limited. In addition to the untimed and timed administration of

word recognition and sentence verification items participants also completed two clusters of PISA reading items. Results revealed that the correlation between the untimed measures of word recognition and sentence verification was only of medium size. However, the correlation between the timed measures was substantially raised and significantly higher. As regards the association with reading comprehension, the untimed measures of word recognition and sentence verification showed correlations of moderate size. Most importantly, the corresponding correlations of the timed measures with reading were significantly higher. Overall, the results suggest that timed administration of items increases convergent validity by eliminating the confounding with the decision on speed. Implications for measurement and generalizability of the proposed testing procedure are discussed.

Strategies within Complex Problem Solving: Inquiries into Exploration Behaviour in MicroFIN

Andre Kretzschmar, Jonas C. Müller and Samuel Greiff
University of Luxembourg, Luxembourg-Kirchberg, Luxembourg
e-mail: andre.kretzschmar@uni.lu

Complex Problem Solving (CPS) is a prominent representative of 21st transversal skills, empirically connected to a broad range of outcomes and recently included in educational large-scale assessments such as PISA. A new measurement instrument of CPS, MicroFIN, combines psychometric quality with more heterogeneous demands in terms of solving the problems used for assessment. Thereby, MicroFIN, offers the opportunity to increase our knowledge of problem solving behaviour in various problem situations (e.g., exploration behaviour). Our Paper explores the consequences of heterogeneous demands within computer-based Complex Problem Solving assessment. With regard to exploration behaviour we examine the applicability and empirical usefulness of different exploration strategies. Based on an established measure of strategic behaviour (VOTAT) and requirements analysis of CPS tasks, we developed a modified conception of an exploration strategy (nested-VOTAT) suitable for a broader range of tasks. We test its utility on an empirical basis (N = 565) using log file data analyses, confirmatory factor analyses, and structural equation modelling. With the help of confirmatory factor analyses we found a separable dimension of exploration behaviour besides the established CPS dimension of knowledge acquisition and knowledge application for MicroFIN. Relations of the exploration dimension to the other two dimensions of MicroFIN as well as to another CPS operationalization (MicroDYN) were as expected and indicated evidence for convergent validity. The present study highlights the potential of process data for the assessment of CPS. Using log file data enables the examination of exploration strategies in a broader range of problem settings and, thus, improves the assessment and the theoretical understanding of CPS.

Behavioral Process Data from Internet-Based Testing

Ulf-Dietrich Reips
University of Konstanz, Konstanz, Germany
e-mail: reips@deusto.es

Since 1994 the World Wide Web has been used in scientific data collection, including testing of participants via the Internet. Any tests and inventories can easily be administered via Internet, the Internet has thus become more and more important for psychological assessment. Behavioral process data, such as so-called paradata, can be collected via the Internet and provide information about participant behavior beyond self-report. The talk will provide an to basic concepts and techniques of collecting and analysing behavioral process data from Internet-based tests and presents suggestions for good practice. The talk will include an up-to-date

overview of techniques, methods, and tools for Internet-based testing, with a focus on behavioral process data. Empirical results from investigations into the validity of several of the behavioral process collection techniques for Internet-based testing will be discussed. It will be demonstrated how to set up and analyze behavioral process data from Internet-based tests with the Scientific LogAnalyzer (Reips & Stieger, 2004) that is available from http://www.sclog.eu. Options for the gathering and analysis of behavioral process data will be compared with those available via public services like Google Analytics and ExtremeTracking and in our own tool Testor (http://testor.pro/). The author further reports on results concerning the correspondence of self-report and behavioral processes in Internet-based testing and measurement. Some recommendations are developed for researchers, authors, reviewers and editors of articles reporting behavioral processing results from Internet-based testing. Finally, an outlook to the future of behavioral process collection techniques for Internet-based testing will be provided.

Benefits of Process-Related Information for Explaining Differences between Administration Modes

Ulf Kroehne, Frank Goldhammer and Carolin Hahnel

German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany

e-mail: kroehne@dipf.de

Before answering questions in a reading test, the text has to be read, presented either on a computer (CBA) or in a printed booklet (PBA). Comparing the results of CBA and PBA for PISA reading assessment, a mode effect concerning item difficulty as a general shift was found in a previous analysis. Moreover, inter-individual differences in behavioral process data had been found for CBA, such as the reading time and the frequency of navigation within units. The study aimed at explaining the observed mode effect by characteristics of the test-taking process, which might provide the causal link between administration mode and the measured competence. Digital pens have been used to gather process-related data from PBA assessment, and indicators were derived from CBA and PBA in a comparable way. In an experimental design, random equivalent groups of 15-year-old students answered one or two clusters of PISA print reading either PBA with digital pens or CBA. In total, data of 851 students have been included. Explanatory IRT-models were used to test hypotheses about the influence of process-related differences between CBA and PBA on the mode effect, simultaneously to the influence of item characteristics and design factors. We found the remaining mode effect to be reduced if the response time of the first item within each unit had been included in the IRT model, although we found no difference concerning navigation related indicators. With this study we extended comparisons between modes to an important characteristic of the test taking process, such as the time individuals spent on reading texts in CBA compared to PBA. Overall, the results suggest that reading time, as chosen by test takers, is an important explanation for the observed mode effect. The theoretical implications of this mediator variable as well as possible further research questions will be discussed.

Do Publications Adequately Acknowledge Permission to Use Others' Tests in Research?

Thomas Oakland
University of Florida, Gainesville, Florida, United States
e-mail: oakland@coe.ufl.edu

Test use in research occurs commonly. Some research relies on author-developed tests while other research relies on tests published in scholarly journals, books, or are commercially published. The contents of scholarly journals and books as well as commercially published tests are copyright protected. Thus, researchers are required to obtain permission to use tests from those who hold copyright to them. This standard seemingly is violated in various countries (Oakland & Iliescu, in press; Oakland et al, 2012) and may be especially overlooked in countries that have a shorter history of psychology and test use (Iliescu, 2012; Iliescu & Vercellino, 2010). This Pannel examines the scope to which researchers obtain and acknowledge permission to use previously published tests, discusses the need to draft a standard that governs these behaviors, and discusses ways such a standard may be both communicated and enforced. Issues will include the rights and responsibilities of test publishers and those engaged in research and publishing.

A Quantitative Segmentation of Stakeholders' Opinions About Legal Test Usage

Dragos Iliescu
SNSPA University, Bucharest, Romania
e-mail: dragos.iliescu@testcentral.ro

Practitioners and researchers are required to obtain permission to use tests from those who hold copyright to them. However, oftentimes this does not happen, especially in countries with only a short history of psychology and test use. Few if any investigations look into the motives why such illegal and unethical behaviors are generalized in some cultures. Common sense prescribes the reasons for such behaviors to a number of motives, among them insufficient knowledge of ethical and legal obligations, no awareness for standards of practice, lack of resources for practitioners, a negative attitude against test publishers on behalf of key stakeholders, and others. The current paper looks at the issue of illegal test usage through the use of Q Methodology. Q Methodology is a combined quantitative-qualitative method, especially useful for opinion and attitude research and segmentation. A Q-Sort of 21 statements was administered to a P-Set of 46 stakeholders from Romania. Among the participants we had students, researchers, practitioners, university professors, test publishers, representatives of public organizations, and test takers. A number of 4 segments of opinion emerged, which are discussed in detail.

The Rights and Responsibilities of Test Publishers and Researchers in Using Tests for Research

Hazel Wheldon
MHS Inc, Toronto, Canada
e-mail: hazel.wheldon@mhs.com

Many of the leading psychological assessments are owned and published by for profit companies. However, these tools are important sources of data for students and academics conducting research studies and when properly translated and adapted for other countries can add valuable insight for professionals and clinicians around the world. Since the test publishers and academics/researchers have competing interests, the challenge is to find a balance between the

needs of the publisher and the needs of the researcher. Throw into this mix the challenges that developing countries have in paying North American rates for psychological assessments and the challenge becomes even greater. In this part of the symposium, we will examine the rights of the test publisher and balance that against the responsibilities that they have to support both academic research and to assist developing countries access some of the most valuable individual data. The right balance is not only beneficial to everyone but important for MHS as an organization in supporting our mission statement of 'improving the lives of individuals and communities around the world'.

Do Publications Adequately Acknowledge Permission to Use Others' Tests in Research?

John Hattie

University of Melbourne, Carlton, Melbourne, Australia

e-mail: jhattie@unimelb.edu.au

Who owns tests? After a test is published in a Journal for the first time, indexed in a compendium (such as Buros Tests in Print, or Goldman's Tests in Journals), placed on a web site, or published by a commercial publisher – who owns the copyright and what are the "standards" that Journal editors should use to ensure compliance. Should journal editors add a standard question ensuring that authors have rightfully sought permission to use any test in the same way many ask about ethics compliance?

Copyright Infringement as Theft and Fraud

Dave Bartram

CEB's SHL Talent Measurement Solutions, Thames Ditton, United Kingdom

e-mail: dave.bartram@shl.com

Copyright infringement is a form of theft and a form of fraud. Not only can it derive people who have invested in the development of materials of the income they need to make good their investment, but also it provides a way for those who infringe copyright to 'pass off' materials as their own when they are not. The development of tests and the adaptation of tests is an expensive and resource intensive process. If it is to be carried out well, it is essential that developers are given the chance to sell their wares without suffering from acts of theft and fraud of their intellectual property. As with any form of anti-social activity, there are two complementary approaches that need to be taken to counter it. First there needs to be an enforced legal system with appropriate penalties for infringement and second those liable to be victims need to take steps to maximise their security.To protect you house against burglary, you do not leave the doors and windows open and rely on the law to stop people! Taking precautions is particularly important in countries where enforcement is lax or where the legal system is not well developed. Security measures that can be taken range from web-patrolling to detect illegal copies of test content, to the use of internal security procedures to protect algorithms and scoring protocols. Paradoxically, the move from paper to online testing has both increased the ease of copyright infringement and increased the ease with which it can be detected.

What Do you Do when your Items Start Appearing in Other People's Tests?

John Rust

The Psychometrics Centre, University of Cambridge, Cambridge, United Kingdom

e-mail: jnr24@cam.ac.uk

In classical psychometrics the unit of sale is very clearly the whole test. However, as we enter the modern era in which adaptive testing is increasingly becoming the vogue, the focus moves to the use of the individual items from which the tests are constituted. The very nature of a classical test limits our ability to alter its items, as any significant changes to these require a re-standardization, often a costly and time-consuming business. But within a modern psychometric Item Response Theory (IRT) framework, test items can be changed, added or removed from a test while in use, so long as item level data is being collected for parameter estimation. For adaptive testing it items and item banks, rather than tests,that require copyright protection. The academic and publishing community has until now been somewhat laid back about this. Particularly within clinical applications, items are often unashamedly 'borrowed' from a variety of tests and recombined at whim, with very little though being given to the enormous amount of work that originally went into the choosing, testing and evaluation of these items by the original author. While this has often been tolerated in the past, guidelines on item use are now urgently required.

How Do we Prepare Psychometric Specialists?

Thomas Oakland

University of Florida, Gainesville, Florida, United States

e-mail: oakland@coe.ufl.edu

Literature on current methods and models for preparing psychometric specialists is sparse. This symposium provides this information from three highly regarded programs. This symposium provides possible benchmarks to assist universities that have psychometric programs as well as universities that express interest in developing or expanding them.

Preparing Psychometric Specialists at the University of Groningen and the Role of the IOPS Organization in the Netherlands

Rob R. Meijer

University of Groningen, Groningen, Netherlands

e-mail: r.r.meijer@rug.nl

In this talk I will present information about the curriculum at the department of psychometrics and statistics in Groningen and I will talk about our experiences with national and international students. Furthermore, I will sketch how we select research master students and how we select students for a Phd to fulfill within a local graduate school. Finally I will talk about IOPS, a succesful research cooperation between different psychometric departments in the Netherlands.

Preparing Doctoral Psychometrics Specialists

Kurt Geisinger
The University of Nebraska-Lincoln, Lincoln, NE, United States
e-mail: kgeisinger@buros.org

This presentation describes the needed skills for psychometricians to work successfully in psychology, education and other related settings. It addresses the psychometric, quantitative/statistical, and psychological coursework as well as the applied experiences including internships that provide students with the opportunities to gain the appropriate knowledge and to advance these skills. This presenter believes that the three foci of any psychometric program need to be on the knowledge of testing methodology, statistical procedures, and the content of the psychological theories and educational programs. All three are needed if an individual is to succeed as a psychometrician. Suggested courses for each of these areas are provided in this presentation. In the area of psychometrics and testing, all psychometricians need to have a basic course taught using a book similar to Anastasi's classic text, one providing information on classical psychometrics, and at least one on item response theory. At Nebraska other courses include courses on equating, validity theory, and testing in various contexts (the schools, high stakes testing) and with various populations (e.g., ethnic and language minorities, immigrants, and those with disabilities). Generalizability, standard setting, test construction techniques, computer-based testing, and fairness are other course topics worth considering. In addition to basic statistics, psychometricians should be competent in ANOVA, multiple regression, multivariate statistics, structural equation modelling, and multi-level modelling. Non-parametric statistics would be another possible and useful topic. Finally, the content of testing is most varied. Today, it could include traditional personality; abilities for schools, employment and/or clinical and neuropsychological services; school achievement; and learning, as mere examples. Finally, it is important that students have active assistantships or internships where they can first observe, then practice their psychometrics. Some experience working within organizations and with groups is very useful. Working under supervision too is quite beneficial to receive feedback on one's approach.

University Training Programs for Specialists in Educational and Psychological Measurement

Ronald Hambleton
University of Massachusetts, Amherst, MA, United States
e-mail: rkh@educ.umass.edu

Over the last 20 years there has been a rapid rise in the need for psychometricians in the United States. In the field of education especially, the 50 states have expanded their testing programs to include more grade levels and more subjects. There is more interest today than there has been in diagnostic testing to support educational programs. In addition, more professions have expanded to include credentialing exams, and the standards for these tests are much higher than they have been with focus on job analyses, validity evidence, equating scores, and standard-setting. And with all of these testing programs, have come new types of uses (criterion-referenced assessment), new forms of test administration (computer- based testing), new test designs (adaptive and multi-stage assessments), new models of testing to support test development, administration, and uses (i.e., item response theory), and new types of questions and machine automated scoring. In this presentation we will describe the doctoral training program at the University of Massachusetts to prepare young professionals to move into the applied and important world of modern measurement. That training is a mix of course work (we teach 22 different courses), cutting edge seminars, and practical experiences that include

internships, and experiences in test development, data analysis, test score equating, standard setting, and more. The importance of specific courses, readings, and training experiences will be addressed during the presentation.

How do we Prepare Psychometric Specialists

John Hattie
University of Melbourne, Carlton, Melbourne, Australia
e-mail: jhattie@unimelb.edu.au

This paper will outline the skills required to prepare psychometrics, and includes • a deep understanding of psychometrics • an understanding of newer models of cognitive task analysis, • procedures in mining of large data sets (such as key stroke analyses), • preparation of reporting using modern human-computer interface methods • have high levels of interpersonal and communication skills, and • policy analyses. All these skills are unlikely in one person but more than one of these skills are needed.

Adapting a Measure of Adaptive Behavior: Challenges Encountered and Lessons Learned

Thomas Oakland
University of Florida, Gainesville, Florida, United States
e-mail: oakland@coe.ufl.edu

Adaptive behavior refers to those practical skills used daily to function and meet the demands of one's environment, including the skills necessary to effectively and independently take care of oneself and to interact with other people. The U.S. developed Adaptive Behavior Assessment System-II is designed to assess adaptive behavior consistent with current theory and clinical guidelines. The ABAS-II has been adapted for use in Denmark, France, Israel, Romania, Republic of China, Spain, Sweden, and for French Canadians and is in press in Czech, Germany, Italy, and the People's Republic of China. The purpose of this symposium is to identify prominent challenges and lessons learned during the adaptation process in Israel, Romania, and Spain.

Challenges Encountered and the Lessons Learned when Adapting the ABAS-II into Hebrew

Dennis Bernstein
PsychTech Ltd, Jerusalem, Israel
e-mail: dbern@psychtech.co.il

The Israel ABAS-II was released in late 2013 in Hebrew and has been welcomed by the professional community. In 2011 PsychTech was approached by the Ministry of Social Services in order to develop a formal translation and validated version of the ABAS. The work began with the development of a Steering Committee composed of private and public sector professionals and additional consultants. An initial translation was completed on all the forms. Several meetings were then devoted to review every item. Most items maintained their integrity without need for cultural or linguistic change. About 20% of the items needed rewording to adapt to Hebrew semantics or terminology. Some of the items needed cultural equivalents. The second stage involved reviewing the items with the test's author to insure transfer of intended meanings. The third step was pretesting in the field and minor adaptations as a result. For the data collection itself, two groups were used. The clinical group (N=76) was drawn from settings where the

diagnoses were already established. The normative sample was collected through households around the country (N=65). Data analysis was conducted to validate the level of Intellectual Disabilities in the clinical groups (mild, moderate, severe) and confirm that the normative groups were scoring within the expected ranges (i.e. 9-11 for the level of Skill Area scales and 85-115 at the level of Adaptive Domain scales). The data analysis found only one scale to be somewhat lower (Self Care), indicating that these items are achieved at a later age within the Israeli context. Since this was a validation study, the decision was made to modulate some of these items downward (i.e. make them easier). This scale was then re-administered to a small additional group (N=19) confirming that the mean rose to within the expected range.

Adaptation of a Measures of Adaptive Behavior in Romania

Dragos Iliescu

SNSPA University, Bucharest, Romania

e-mail: dragos.iliescu@testcentral.ro

The present paper describes the adaptation process for the Adaptive Behavior Assessment System, Second Edition (ABAS-II) to the Romanian culture. Steps taken to insure a valid translation and cultural adaptation process are described. All three forms of the assessment system are discussed (Teacher, Parent, and Self). Data from more than 4000 cases who completed the Romanian ABAS-II are examined focusing on internal consistency, intercorrelations, and factor structure equivalence. Multitrait-Multimethod convergence of the three forms is also examined. Finally, the ABAS-II diagnostic validity for a number of diagnoses is examined, as well as its incremental validity over measures of cognitive ability.

ABAS-II Spanish Adaptation: Challenges and Solutions

Irene Fernández Pinto

TEA Ediciones, Madrid, Spain

e-mail: irene.fernandez@teaediciones.com

Adapting a test from a culture to another is always a great challenge since it requires the adapted instrument to be both equivalent to the original in terms of the construct measured, and adequate to the target culture. Sometimes both requirements are difficult to attain simultaneously and creative solutions are needed. In this presentation we intend to analyze the specific challenges in the adaptation of the Adaptive Behavior Assessment System – II to the Spanish speaking population. This process not only involved a linguistic adaptation of the items, but also a cultural adaptation of the specific behaviors that allow us to identify the degree in which a person functions successfully in his or her environment in the target culture. In a test like ABAS-II, the behaviors measured need to be specific, and thus, they are often culturally charged. For this reason, both linguistic and cultural adaptation were particularly complex to implement, especially taking into account that the adapted test was intended to be administered not only in Spain but also in a wide range of American Spanish Speaking countries. Thus, the solutions given to this problem will be presented. The standardization process, that requires collecting a wide and representative sample of the target population, regarding various sociodemographical variables, will be described, as well as the psychometrical properties of the Spanish adaptation of the ABAS-II. The results show that the adapted test constitutes a reliable and valid measure of adaptive behavior in the target population, and its items include contents both easy to understand and culturally relevant.

Innovations in the Analysis of Ratings and Raters

Edward Wolfe
Pearson, Iowa City, IA, United States
e-mail: ed.wolfe@pearson.com

Ratings are utilized in many measurement settings, worldwide, and a primary concern of those who make decisions based on ratings is the quality of the resulting measures due to the inherent subjective nature of the rating process. In order to safeguard against measurement error that may be introduced into these measures, numerous processes are commonly employed to monitor the quality of ratings both during and at the conclusion of operational scoring projects. This symposium brings together four papers that describe innovations in the analysis of ratings and raters. The first paper (Arce-Ferrer, et.al.) presents new procedures for monitoring raters in large-scale scoring environments based on several indices of rater quality. The second paper (Cohen) demonstrates how rater quality indices may be biased in the presence of ignored dependence between ratings. The third paper (Wind, et.al.) describes how rater effects influence the accuracy of scores assigned by automated scoring engines. The final paper (Keller, et.al.) seeks to determine the impact of regression-based corrections that are applied to centre-level scores by a monitoring agency. The four papers provide insights from scholars who are experts in the analysis of ratings, and each paper addresses an innovative process for monitoring raters. The four papers adopt diverse perspectives, and the discussant for the symposium is an expert with years of experience in both the testing industry and academia. With the continued popularity of the use of ratings in high-stakes assessment contexts, the importance of monitoring raters is self-evident. In order to confirm that ratings are of the highest quality possible and are, therefore, as free from measurement error as possible, effective and efficient rater monitoring procedures are paramount. This symposium presents new insights from leading scholars into those processes.

Towards an Approach for Monitoring Raters and Ratings in Large-Scale Scoring Environments

Alvaro Arce-Ferrer
Pearson, San Antonio, United States
e-mail: alvaro.arce-ferrer@pearson.com

Monitoring the quality of scoring performance tasks is far more complicated than that of scoring multiple choice answers. Increasing the standardization and objectivity of human scoring requires use of a clear and unambiguous benchmark, thorough training of raters, and careful monitoring procedures to evaluate the quality of each rater. The rater is expected to rate accurately and consistently over time. In operational settings, summative evaluations of rating quality are often carried out with multiple manifest indices such as rater agreement, rater error and systematic bias, and rater accuracy. The relationship between indices of rater quality was studied to formulate an approach within the statistical process control framework that might yield formative and summative evaluations of rating quality. The study used operational data gathered from the writing task of the PET (Psychometric Entrance Test) used for higher education admissions in Israel. Our approach presents three layers of inspection of rating quality: individual, batch, and process. At the individual level, ratings for an individual rater are compared to other raters: a negative difference between an individual and his/her group indicates severity and a positive difference indicates leniency. At the batch level, statistical measures of each rater are summarized to construct an overall test of significance on raters'

consistency for a given batch. At the process level, the batch level statistics are combined to provide an overall test of significance with regard to raters' consistency for the scoring process. This third layer inspects presence of drift in, and abrupt shifts between, batches of ratings. Preliminary analyses of raters' quality using traditional approaches have resulted in low correlations between indices. To date, our approach has been evaluated using logical and statistical arguments, pending its evaluation based on empirical data, which is planned for winter 2014.

Influence of Rater Effects on the Training of Automated Scoring Engines

Stefanie Wind[1], Edward W. Wolfe[2], George Engelhard[3], Mark Rosenstein[4] and Peter Foltz[4]

[1]Emory University, Atlanta, United States; [2]Pearson, Iowa City, United States; [3]The University of Georgia, Athens, United States; [4]Pearson, Boulder, United States

e-mail: swind@emory.edu

Training sets for automated scoring engines (ASEs) are typically based on human-assigned ratings (HARs), that are regressed onto response features to determine which features constitute the best set of human-assigned score predictors. Because HARs are used to define the scoring model for the ASE, it is essential to consider the quality of ratings in ASE training sets. Previous research has considered the influence of training set characteristics related to the number of essays, the distribution of score points, and the level of human rater agreement and reliability; however the influence of rater effects based on latent trait models on the quality of ASE ratings has not yet been considered. This study evaluates the impact of rater effects in HARs on the training and subsequent accuracy of ASE ratings. Data came from a large-scale rater-mediated middle school writing assessment in which essay responses of 400 students were rated by 131 raters. In addition, a Pannel of experts assigned consensus true scores. HARs were examined for evidence of three rater effects: severity, centrality, and inaccuracy. ASE training sets based on raters evidencing these effects were compiled and an ASE was trained on a portion of the essays. The remaining essays were rated by the ASE and performance of the ratings assessed for each training set. Initial analyses indicate a wide range of rater effects in the HARs. Calibration of the engine demonstrates that ASE training is sensitive to several of these effects but that their detrimental influence on ASE accuracy can be reduced. As automated scoring technology use increases, it is essential that we find ways to optimize the training of ASEs. This study provides insights into how training set composition influences that process.

Quality Control in Essay Rating: Estimating the Reliability of Essay Ratings in Cases of Rater Arbitration

Yoav Cohen

NITE (National Institute for Testing & Evaluation), Jerusalem, Israel

e-mail: yoav@nite.org.il

A common method of quality control in essay rating is the use of a third rater whenever the ratings given by two assigned raters differ by more than a predetermined threshold. In some high-stakes testing programs, the final rating is the sum of the third with the rating which is closest to it. The purpose of this study is to show that the inter-rater reliability estimate under these conditions is a biased estimate of the true reliability. We assume that essay ratings conform to Classical Test Theory, i.e., that the observed rating is the sum of a true rating and an independently distributed error term. The first part of the study is based on simulations in the framework of CTT, and the second part is an empirical investigation of the same research question. The empirical investigation is based on the estimation of true ratings. We estimate the

"true rating" by averaging a relatively large number of ratings per essay. Two groups of 13-15 trained raters were asked to rate a total of 500 essays. Thus, for each essay we have both observed ratings and "true" ratings, and thus we can estimate better the reliability of the ratings. We found good agreement between the results of the simulations and the empirical data. Both the simulations and the empirical data show clearly that estimating reliability by correlating two ratings per essay is a biased estimate of the true reliability, and that this is due to what I termed "The Third Rater Fallacy". We show that the estimate of rating reliability, when calculated by the correlation between two ratings after an arbitration process, is biased and should not be used. Moreover, we show that certain methods of rater arbitration result in an increased error of measurement.

What is Considered Close enough? A Consideration of External Moderation for Complex Assessment Tasks

Lisa Keller[1], Joseph Rios[1] and Edward Wolfe[2]

[1]University of Massachusetts Amherst, MA, United States; [2]Pearson Educational Measurement, Iowa City, United States

e-mail: lkeller@educ.umass.edu

Scoring open-ended assessment tasks is a potentially expensive endeavor; having teachers score student responses to these tasks could result in two benefits—reduce the costs associated with obtaining scores while increasing the likelihood that those scores can be used to inform instruction in a timely manner. However, teacher-assigned scores that are used for decisions external to the classroom should be held to the same standard as scores assigned by testing agencies. One method for monitoring local rating effects is to have a sample of student responses rescored by a testing agency; a correction can be applied when local scores deviate too much from the agency scores—a process called moderation. The goal of this study is to determine the impact of varying rules for the application of moderation decisions on measures of student performance. A simulation study was conducted based on data from an operational assessment that uses a moderation process to examine the effects of varying rules regarding how much deviation from testing agency scores is tolerated with respect to the impact of the correction on scores. Preliminary results suggest that reliability, bias and sample size each have a predictable impact on the accuracy of moderated scores. Moderation processes are utilized in various forms of internal assessment in the United Kingdom. In the United States, assessment consortia are moving toward internal assessment models. Hence, external moderation is a potentially useful method for making the scoring of assessments both more efficient and accurate. Based on simulation study results, we are offer recommendations for how to choose appropriate tolerance intervals and sample sizes.

Assessing Candidates with Disabilities - Current Practices, Challenges and Future Direction

Tanya Yao

CEB- SHL Talent Measurement, Surrey, United Kingdom

e-mail: tanya.yao@shl.com

Recently, the area of assessing disabled candidates has been increasingly popular. Regulations in countries such as UK/US have stressed the importance of inclusion. Therefore it is at the interests of test publishers, recruiters, employers and practitioners to use testing products that are

inclusive of disabled candidates. In this symposium, the area of assessing disabled candidates will be approached from a range of perspectives. To make testing inclusive for disabled candidates, it is most important to provide them with reasonable adjustments during assessments. First presentation will illustrate what reasonable adjustments are and discuss difficulties and research needs in providing reasonable adjustments and interpreting results. Second presentation will discuss the challenges and opportunities faced from test publisher prospective. It opens up an 'industry research agenda' in this area and will also share experiences as test publisher, both from practice perspective as well as research perspective. Assessment Center is a common method many recruiters use. The third presentation will discuss disability accommodations at Assessment Centres, focusing on the difficulties disabled candidates might face, and the impact on other candidates. The experts will use research as well as case studies and testimonials from disabled candidates to tell us how; this topic will be exposed on the fourth presentation. Disability confidence is vital to provide suitable adjustments and assist them through tests. The barriers faced by disabled candidates and how we can plan for these and overcome them will be discussed. Although an answer could not be provided here, we aim at raising awareness of this topic. We hope the profession will be able to provide some answers to these questions such as validity and reliability of accommodated assessments and nature of accommodations requested. By assessing these areas, we hope to be able Promoting the assessment of disabled talent in selection and development.

Disability Accommodations at Assessment Centres

Helen Baron

Independent, London, United Kingdom

e-mail: helen@hbaron.co.uk

Assessment Centres are commonly used in occupational and organisational assessment and involve the concurrent assessment of two or more individuals using multiple exercises and assessors with at least some interaction between participants. The British Psychological Society has convened a working group to develop standards of practice for Assessment Centres. This paper will discuss some of the proposed standards and guidance with respect to assessing people with disabilities. There are a number of issues. Firstly there are a variety of tasks and exercises for which the candidate may need some accommodation. The tasks are likely to be designed to use a variety of information input media, processing skills and output methods. Therefore there is a high likelihood that a person with a disability will need some accommodation to be able to demonstrate their capability effectively. Secondly the length and physical demands of the event generate their own issues for people with disabilities. The need to move around a large unfamiliar venue, concentrate for extended periods and even the timing of breaks and opportunities to eat and drink can all raise issues that need to be addressed. Lastly assessment centres involve the assessment of a group of people together. Participants typically interact with each other in at least some of the exercises. Therefore consideration is required as to how the presence of the disabled participant affects all members of the group and the whole assessment process. For example asking other participants to accommodate a hard of hearing candidate by avoiding speaking over each other and ensuring their mouth can be seen when speaking might change the dynamics of a group exercise for everyone and therefore their assessment results.

Assessing Candidates with Disabilities - Current Practices, Challenges and Future Direction

Kate Headley
The Clear Company Ltd, Norley, United Kingdom
e-mail: kate.headley@theclearcompany.co.uk

How do employers ensure that all tests are administered fairly when some participants have a variety of disabilities or long term health conditions, which can have an adverse impact on their ability to undertake tests? The problem facing test providers and employers is gaining a sufficient understanding of the impact of an individual's disability or health condition and, therefore, the possible effect it will have on the outcome of tests. In addition, test providers and employers need to be certain what adjustments can be made and what potential impact there is on the efficacy of tests. To show how employers can widen their available talent pool and gain a competitive advantage in the war for the attracting skills by getting this right. Methodology: We will utilise real case studies from disabled people and employers showing how small changes to an assessment process can have a significant positive impact on outcomes. Our research will show how asking the right questions at the right time engenders trust, allowing every candidate to be assessed fairly but that doesn't mean 'lowering the bar' for skills requirements. In fact, with a wider pool of talent to select from means selection criteria can be adhered to tightly. Test providers and employers will see: • The importance of planning for adjustments • How to create an environment of trust by asking the right questions in the right way • How to develop recruiter (assessor) and candidate (employee) confidence • The danger of having inaccessible testing platforms • How getting this right will improve the quality of hire (employee development) decision • That this is not about lowering the bar. Making adjustments to assessment process should be a mainstream feature of the assessment planning and implementation process.

Increasing Inclusivity in Psychometric Assessments through Reasonable Adjustments

Tanya Yao
CEB- SHL Talent Measurement, Surrey, United Kingdom
e-mail: tanya.yao@shl.com

The UK Equality Act states that all employers have a duty to provide reasonable adjustments to facilitate optimal performance in their employees with disabilities. Talent management organisations providing psychometric assessments have a legal obligation to support their clients in providing adjustments to disabled candidates who wish to complete these assessments. While the concept of adjustments, e.g. additional time in exam, has been around for some time, little research has been conducted on the effect of adjustments on psychometric assessments. The aim of this presentation is to provide an overview of what has been done and what is yet to be done to ensure fairness throughout the assessment process through reasonable adjustments. Current practice Different adjustments are available depending on types of disability and needs of candidates. Additional time is often used for candidates with dyslexia. Font-size and line spacing can be increased for those who are visually impaired. Braille or alternative colour versions are also available. Online assessment delivery platforms can be made compatible with screen readers. Assessments using sound may have subtitles for those who have hearing impairments. Research Potentials Anxiety or lack of knowledge of relevant legislation and disabilities may result in uncertainty for recruiters about how to approach adjustments. There is a vast array of disabilities and possible adjustments, and there remains a stigma attached to disability such that disclosure does not always occur. Additionally, once a suitable adjustment has been implemented, there is the question of how to interpret the assessment score. Future Directions More research is needed on the types of adjustments available, and their

validity. Research on the extent to which test scores vary as a function of disability and the candidate's experience of online psychometric assessments are also important. Future assessment development may be influenced by the results of this research.

Research Challenges and Opportunities in the Assessment of Disabled Talent in Employee Selection and Development

Rachel Owens

CEB SHL Talent Measurement, Thames-Ditton, United Kingdom

e-mail: rachel.owens@executiveboard.com

The Equality Act in 2011 means that recruiters and test publishers have an obligation to ensure that reasonable adjustments are provided to disabled candidates in order to ensure optimal performance. We discuss the opportunities and challenges that we have found to be associated with the assessment of disabled-talent, and in doing so hope to encourage more interest and research in this area. We see research on group differences in psychometric testing as fundamental. This can inform us of the magnitude of score differences and the extent to which there is potential for adverse impact against disabled candidates. It may also inform us of the most effective types of adjustments. Research on candidate experience of online psychometric assessments is also important. This research will contribute to future assessment development, as well as clarifying what adjustments are frequently required or considered most useful. Actions Data is being collected on user experience of our online psychometric assessments as well as the disability types. We are also improving all of our technology and platforms to make the assessments more accessible to people with disabilities. Challenges Disabled-Talent is an area that has received little attention, be it from the perspective of the disabled candidate or from an employer, meaning we have little to go on. Perhaps this is due to the large number of disabilities that exist, or the seemingly endless number of possible adjustments that could be made. The stigma surrounding disability may also be affecting disclosure rates. As an international company, we are further faced with a great deal of cultural and legislative differences between countries which makes creating a standard inclusion strategy difficult. Future Directions We outline our research agenda on Disabled-Talent which includes collaborations with large firms and charities.

Developments in Item Analysis, Latent Variable Methods, and Scoring

Bruno Zumbo

University of British Columbia, Vancouver, Canada

e-mail: bruno.zumbo@ubc.ca

The psychometric properties of item responses and scoring are essential issues in measurement and test development. For example, despite the increased use of item response theory (IRT) methods with non-cognitive measures, little research has been conducted to further our understanding of the substantive meaning and interpretation of item parameters for psychological measures; and when an item is flagged as a DIF item for different versions of language, how can we know if the item functions differently due to the translation rather than other factors. The purpose of this symposium is to present recent developments on these issues. The papers in this symposium cover a variety of topics on item responding, differential item functioning (DIF), and scoring. More specifically, these topics include: (a) an investigation into the psychological meaning of IRT item parameters, (b) when and how to use propensity score

methods to our advantage in DIF analysis, (c) lessons learned in applying DIF methods to the Youth Self-Report Scale, and (d) a comparison of observed-score and latent variable test (scale) scoring methods. This symposium serves as a window into recent developments of value to practitioners and researchers alike.

Logistic Regression Differential Item Functioning Analysis Using a Propensity Score Approach

Yan Liu[1], Bruno D. Zumbo[2], Paul Gustafson[2], Edward Kroc[2], Yi Huang[2] and Amery D. Wu[2]
[1]Harvard University, Boston, MA, United States; [2]University of British Columbia, Vancouver, Canada
e-mail: liuyanspa@gmail.com

Conventional DIF methods such as logistic regression, Mantel-Haenszel, or IRT methods can only flag DIF items but cannot, on their own, sort out the causes of DIF. Therefore, it is important for researchers to use more elaborated DIF analysis methods to examine DIF items after purifying other confounding variables. In some contexts, propensity score matching can be used to balance the characteristics of non-equivalent groups, so that the two groups become comparable. The purpose of the present study is to demonstrate how to use propensity score methods in logistic regression analysis for examining DIF due to the adaptation and translation, in our case from English to French. The data were retrieved from Trends in International Mathematics and Science Study (TIMSS) 2007. Booklet 1 of the TIMSS 2007 Grade-8 mathematics test was used in the. Twelve variables were chosen from students' background questionnaire, which were used for propensity score estimation. A sample of 820 students from Canada were selected, with 542 (65.9%) English and 281 (34.1%) French speakers. The DIF analyses were conducted using both conventional logistical regression as well as logistic regression based on a modified propensity score approach. The results showed that some of the DIF items flagged by the conventional logistic regression were shown to not be so when using the propensity score approach. The findings indicate that the adaptation and translation did not cause DIF for some of the DIF items flagged by the conventional DIF method and test takers should have the same probability to get the item correct after controlling for their background variables. This new approach may help test administrators to make a more accurate decision about keeping or removing items for a test.

From Measurement Models to Scoring Methods: An Application to Group Differences

Ze Wang, Steven Osterlind, Wendy Reinke and Melissa Stormont
University of Missouri, Columbia, United States
e-mail: wangze@missouri.edu

For each measurement model, there is a consistent scoring method. The use of non-intended scoring methods could result in different about the same substantive research hypotheses due to the statistical assumptions associated with the corresponding measurement models. 1) To compare four measurement models and their scoring assumptions; and 2) to show that different scoring methods could result in different in terms of statistically significant group differences and in terms of the magnitude (i.e., effect sizes) of those differences. The four measurement models are: 1) classical test theory model, 2) first-order CFA with continuous indicators, 3) first-order CFA with ordered-categorical indicators, and 4) multidimensional graded response IRT model. The four measurement models were imposed on the same data collected from 577 K-3rd grade students in the US. Specifically, their teachers rated these students using the 21-item Teacher Observation of Classroom Adaptation Checklist (TOCA-C). Consistent with the four measurement models/scoring methods, gender differences and differences between

free/reduced lunch status groups in the three TOCA-C subscales were examined. In addition, measurement invariance was tested for the second, third, and fourth methods before latent factor means were compared. For the first model/method, observed group mean differences were examined. The second, third, and fourth methods demonstrated different levels of measurement invariance. Group mean differences were examined with the most appropriate level. Gender differences were statistically significant and their effect sizes were close to medium in the TOCA-C subscales using all four scoring methods. However, for free/reduced lunch status group mean differences, depended on the scoring method. Scoring methods matter for substantive research. Scoring recommendations made for researcher-developed measures should be consistent with the measurement models used at the development stage of the measures. One future research direction is to develop practical tools to ensure such consistency.

An Examination of Differential Item Functioning on Items of the Youth Self-Report

Xinya Liang and Yanyun Yang
Florida State University, Tallahassee, FL, United States;
e-mail: xl07e@fsu.edu

The Youth Self-Report (YSR; Achenbach & Edelbrock, 1991) is a widely used scale for measuring behavioral and emotional problems of youths in 11-18 years old. It consists of 119 items rated on three categories: 0 (not true), 1 (somewhat true), and 2 (often true). Ivanova et al. (2007) found that the initial 8-syndrome structure was consistently supported by data from 23 countries. However, differential item functioning (DIF) for YSR items has rarely been investigated. An item demonstrates DIF if examinees in different groups have unequal probabilities of endorsing a particular category, conditional on their ability levels. This study examined sex and age DIF for YSR items using Generalized Mantel-Haenszel (GMH; Mantel & Haenszel, 1959) and standardized mean difference (SMD; Dorans, Schmitt, & Bleistein, 1992). Data were collected from 1370 youths in 12-18 years old in China. Eighty nine items measuring the following 8 syndromes were included in the analysis: Anxious, Withdrawn, Somatic Complaints, Social Problems, Thought Problems, Attention Problems, Rule-Breaking Behavior, and Aggressive Behavior. For DIF analyses, Male and older (aged 15-18) groups were chosen as reference groups. In GMH analysis, the total raw score was the ability matching variable stratified into 10 levels. The association of groups and responses was tested at .05 level. In addition, the effect size for SMD was calculated. Items were classified into one of the three categories based on National Assessment of Education Progress guidelines: negligible, moderate, and large. Preliminary analysis shows that 27 and 17 items demonstrated moderate to large DIF for sex and age, respectively. Females were more likely to endorse Anxious items but less likely to endorse Rule-Breaking and Aggressive Behavior items. Younger teens tended to endorse Attention and Aggressive items, while older teens endorse Thought, Social, and Somatic items.

Making Sense of IRT Parameters in Non-cognitive Measures by Investigating their Relationships with Five Social-Psychological Factors of Item Responding

Amery Wu
The University of British Columbia, Vancouver, Canada
e-mail: ameryw@yahoo.com

Despite the increased use of item response theory (IRT) methods with non-cognitive measures, little research has been conducted to further our understanding of the substantive meanings behind IRT parameters for psychological measures. The purpose of the present study was to make sense of the IRT item parameters, beyond its mathematical meaning, by examining their

relationships with five social-psychological factors that might influence the process of item responding. The five factors were, wording specificity, availability heuristic, emotional comfort, meaning clarity, and social desirability. IRT parameters were obtained from the responses of 729 adults to the 30 items of the Geriatric Depression Scale (GDS; Yesavage et al., 1983). After testing five alternative models, a 3 parameter logistic model with a-, b-, and d-parameters (i.e., discrimination, difficulty, upper asymptote) showed best fit to the data. Another sample of 30 adults rated the influence of five social-psychological factors immediately after responding to each one of the GDS items. Due to the ordinal nature of the ratings as well as the dependence of 30 item ratings within each rater, the relationship between each of the a-, b-, and d-parameters with the social-psychological ratings were analyzed using multilevel ordinal logistic regression. In this study, we found the a-parameter to be negatively related to the availability heuristic and social desirability. The b-parameter was found to be negatively associated with wording specificity, meaning clarity, and social desirability. The d-parameter was found to be positively associated with social desirability. Only the ratings of emotional comfort in responding to an item were found to be unrelated to any of the a, b, or d parameters. The findings of this study expand our understanding of the substantive meanings behind IRT parameters for non-cognitive measures.

Round Tables & Pannel Sessions

A Decade of Test Security: The Road Traveled, The Road Ahead (Pannel Session)

Steve Addicott[1] and Philip Dickinson[2]

[1]Caveon Test Security, Midvale, United States; [2]National Council of State Boards of Nursing, United States

e-mail: steve.addicott@caveon.com

In the past decade, we've witnessed incredible changes in test security—both good and bad. While new technologies have created ever-greater risks to reliable test results, other innovations have empowered test program leaders with new tools to better protect them. All the while, the stakes in international testing march ever-higher. And, we know test security challenges will not only be different, but in many ways, even tougher. In order to keep up, we'll need creativity and technology to invent new methods and tactics in the ongoing battle for test results that matter. This session will explore answers to the question: what will test security look like in the coming years? Join two test security veterans for a quick look back, and a glimpse into the "crystal ball" of the future of test security. This session will explore illuminating topics that directly impact the validity of high stakes test scores, such as: • As an industry, where have we made the biggest strides in protecting our exams? Where have we lagged? • Looking ahead, how will security risks evolve? How can we protect our programs from these risks? • Which new technologies will prove most damaging, and most helpful? • Most importantly, what should our future test security plans include that they may not currently? Presenters include Phil Dickison, Chief Examinations Officer of the National Council of State Boards of Nursing and Steve Addicott, Senior Vice President of Caveon Test Security. They will not only share valuable "lessons learned" through pivotal test security experiences, but will explore the trends impacting assessment security over the coming years. In doing so, attendees will be armed to thoughtfully improve their security planning now; consider the impact of future threats on their programs; and formulate and implement creative new methods to manage against them.

Practical Applications of Computerized Adaptive Testing (Round Table)

Joe Betts[1], Ada Woo[2] and Anthony Zara[3]

[1]Pearson Vue, Chicago, United States; [2]National Council of State Boards of Nursing (NCSBN), Chicago, United States; [3]Pearson Vue, Bloomington, United States

e-mail: joe.betts@pearson.com

For this roundtable, experts Ada Woo, PhD, Associate Director, Measurement & Testing from National Council State Boards of Nursing (NCSBN), and Tony Zara, PhD, Vice President, Global Credentialing Solutions from Pearson, will meet with interested conference attendees about practical considerations when implementing a CAT program. In the late 1980's NCSBN began researching CAT and in 1994, NCSBN became the first health care organization to utilize computerized adaptive testing for the purpose of licensure. At least one large scale testing organization has moved away from item level CAT because of the large number of calibrated items required to target the full ability range of examinees. Our experts will lead a discussion of this perspective and be prepared to talk about how NCSBN meets the challenge of building and maintaining a large item bank. Computerized testing offers a variety of modes of test administration other than CAT including linear, LOFT, and Multi-stage testing. The experts will lead the roundtable discussion on the merits of item level CAT, the types of research that NCSBN has conducted related to CAT, and the implications for content development and psychometric practices if an organization chooses to utilize a CAT administration.

Copyright: How can we Balance the Needs of Authors, Publishers, Users, Researchers and Clients (Pannel Session)

Ian Florence[1], Juergen Hogrefe[2], Dragos Iliescu[3] and Natasa Kö

[1]Only Connect Limited, Henley-on-Thames, United Kingdom; [2]Hogrefe, United States; [3]SNSPA University, Bucharest, Romania

e-mail: ian.florance@btinternet.com

The international reach of on-line products and services has challenged traditional views of copyright and stimulated new approaches to intellectual property and copyright such as the Creative Commons Movement, as well as creating new ways of using and exploiting what previously had been very protected materials. The experience of consumer entertainment industries are relevant here. This is an area in transition. These are critical issues within the international testing community. How do we balance the needs of public health and social services which are often seeing budget cuts with the needs of commercial publishers to make a profit and invest in new products? Tests are expensive to produce but are developers using this to overcharge ? How can we ensure that innovative assessment research is funded ? Can we agree on the need to end abuses of copyright material, particularly the outright theft of tests in emerging markets ? How do these affect clients, patients and others ? What aspects of tests are actually copyrightable ? This Pannel discussion session involves a group of individuals who are both at the centre of the European test publishing industry as well as having psychological training. They combine experience of both protecting copyright and of markets where the use of illegal material affects everyone involved. In addition to a Pannel discussion the group will welcome challenging questions from all areas of the international testing community.

Broadening Language Assessment Horizons through a Systems Based Approach (Round Table)

Jesse Markow and Timothy Boals

WIDA at the University of Wisconsin-Madison, Madison, United States

e-mail: markow@wisc.edu

This roundtable session will provide an overview of the WIDA language standards and assessments system, including the rationale for key developments to-date and the vision for the next decade and beyond both within the US and in international contexts. World-Class Instructional Design and Assessment (WIDA) has developed a standards and assessment systems approach to language education for English language learners (ELLs) unique in US education. It can also provide insights for teaching and assessing language growth in similar contexts in other countries, especially in K–12 schools where language and content instruction and assessment are both important. Our mission is to advance academic language development and academic achievement for linguistically diverse students through high quality standards, assessments, research, and professional development. Housed at the University of Wisconsin-Madison, as part of the Wisconsin Center for Education Research, WIDA's 33 member states represent over 1.4 million ELLs in Kindergarten through grade 12 in US schools. WIDA partners with the Center for Applied Linguistics in Washington, DC, for language assessment development. Our reach has touched India, Russia, Taiwan, and Qatar. A version of WIDA's language standards contextualized for schools outside the US is available and WIDA's MODEL language assessment is used by international schools with English based curricula. We attribute our rapid growth to the mission and vision of WIDA to develop, enhance, and validate a

comprehensive standards-referenced assessment system for English language learners (ELLs). At this session WIDA will: 1. Share an expanding vision of language assessment internal and external to classroom instruction from the perspective of a multi-state consortium with international reach 2. Describe how language development standards become the genesis for language assessment across contexts 3. Elaborate on a new generation of large-scale, technology-driven language assessments 4. Discuss how interim language measures complement large-scale efforts by providing actionable diagnostic information.

Building an Innovative Test (Round Table)

Jason Schwartz[1], Betty Bergstrom[1] and Philip Dickison[2]

[1]Pearson, Chicago, United States; [3]National Council of State Boards of Nursing (NCSBN), Chicago, United States

e-mail: jsmathematics@gmail.com

This roundtable will provide an opportunity for interested conference attendees to meet and discuss adding innovation into existing tests, or building new tests in which innovation is the keystone. Philip Dickison, PhD, RN, Chief Officer, Examinations, National Council State Boards of Nursing (NCSBN), and Betty Bergstrom, Vice President, Testing Services, Pearson, will lead the discussion based on their experiences over that last two years. Dr. Dickision is currently leading the NCSBN charge to review their current examination and explore innovative item types and innovative test design. Dr. Bergstrom has been at the lead in research related to implementing technology enhanced items. Roundtable participants will have an opportunity to discuss issues related to research on, development of, and utilization of innovation for testing. Adding innovative items, multi-media and special effects can enhance the authenticity of an examination. It can also create a number of possible disadvantages, e.g. add construct irrelevant variability, create issues for examinees with disabilities, add significantly to the cost, etc. Our experts will discuss the pros and cons of adding innovation into an examination along with some insights from their experiences with respect to utilizing technology to enhance the types of items that might be utilized in an innovative manner. Roundtable attendees will have an opportunity to discuss cutting-edge issues related to building an innovative test. Additionally, NCSBN and Pearson, have embarked on a number of studies related to practical content development, scalability, and validity of innovative items. Attendees will have an opportunity to learn about this research and some of the possible future directions that research and development might like to aim. This roundtable would provide an opportunity for other attendees to share their research ideas and hear how NCSBN is approaching their future.

Single Papers

Construct Equivalence using Structural Equation Modeling: Effects of Sample Size Ratios in Multi-group Comparisons

Mathijs Affourtit[1] and Ilke Inceoglu[2]

[1]CEB - SHL Talent Measurement Solutions, Thames Ditton, United Kingdom; [2]Surrey Business School, University of Surrey, Guildford, United Kingdom

e-mail: mathijs.affourtit@shl.com

Structural equation modelling (SEM) is often used to examine measurement equivalence across groups in multi-construct personality instruments. Previous studies have shown that fit indices of single sample models can be affected by sample size, but that this effect is reduced when samples are larger (Iacobucci, 2010), suggesting that using larger samples reduces bias. It is unclear however, whether this can be extended to multi-group comparisons as used for measurement equivalence research, and whether the same recommendations apply. The aim of this study is to examine how model fit in multi-group comparisons is affected by the sample sizes of the individual samples in terms of ratio and absolute sample sizes. Simulated data was used to explore the effects of variations in sample sizes on fit indices. Sample sizes were varied from 100 to 20'000 and in total 10'000 estimations were carried out. Construct equivalence (configural model) was tested by comparing covariance matrices across the two samples, as described in Bentler (1995) and Byrne (2006). This method allows for testing construct equivalence when there is no factorial model defined in the personality instrument used. Good fit indicates that the relationships of the constructs are similar across the tested samples, thereby providing evidence for construct equivalence. Fit indices are plotted by keeping the reference sample size the same and varying the size of the target language. Graphs show V-shapes where lowest fit is found when sample sizes are equal. The results reveal the surprising and important finding, that unbalanced samples sizes can substantially influence fit indices by inflating them. This is an important finding for researchers exploring the equivalence of measures across cultures using a multi-group structural equation model. Results highlight the importance of taking into account the size of the samples that are being compared.

Latent Class Analysis of Large Scale Data from the Multiple Cognitive Abilities

Khaleel A. Al-Harbi[1] and Dimiter Dimitrov[2]

[1]National Center for Assessment, Riyadh, Saudi Arabia; [2]George Mason University, Fairfax, VA, United States

e-mail: khaleel@qiyas.org

The Multiple Cognitive Abilities Assessment (MCAA)is used to identify gifted and talented students in grades 3-12 in Saudi Arabia. The MCAA consists of four dimensions, Verbal Reasoning and Reading Comprehension (VRRC), Mental Flexibility (MF), Mathematical and Spatial Reasoning (MSR), and Science and Mechanical Reasoning (SMR). The goal is to identify latent classes of students across performance on VRCC, MF, MSR, and SMR and the effect of gender on such classes. Latent class analysis (LCA) is used to identify latent (hidden) variables which represent subpopulations of subjects that are not known in advance but, instead, are inferred from the data (e.g., McLachlan & Peel, 2000; Muthén, 2001). The LCA is applied with data for 40,324 students from grades 3, 6, and 9. The role of gender in each latent class is investigated. The LCA was conducted here using the computer program Mplus (Muthén & Muthén, 2010). The decision on how many latent classes to retain was guided by the Bayesian Information Criterion adjusted for sample size, Adj.BIC, the Vuong-Lo-Mendel-Rubin Likelihood Ratio test, and the Lo-Mendel-Rubin Adjusted Likelihood Ratio test. The results led to the identification of five latent classes of students at each grade level. For the study sample (N = 40,324), the latent classes varied in size from (a) 4.6% to 31.8% for grade 3, (b) 9.0% to 37.1 % for grade 6, and (c) 1.7%

to 38.7% for grade 9. Consistently, the highest performance across latent classes was on the VRRC and the lowest on the SMR. Overall males did better than females across all latent classes and grades. The findings in this study, coupled with additional examination of characteristics of the students in each latent class, will provide valuable contribution to the research on gifted and talented students in Saudi Arabia.

Quality Control for Scoring Continuously Administered Tests

Avi Allalouf, Tony Gutentag, Michal Baumer and Marina Fronton
NITE, Jerusalem, Israel
e-mail: avi@nite.org.il

Quality control (QC) procedures for scoring can be divided into two types: QC on tests that are administered to large population groups on a number of set dates using a set test form, and QC on tests administered to small population groups on many administration dates using a variety of test forms. This research relates to the latter, is a highly relevant type of QC, given that the world of testing is headed in this direction. This study used data pertaining to 23,548 examinees who took an online test during four years, to develop QC measures for this test and for continuously administered tests. The study first looked at test delivery: A descriptive examination of weighted sampling of test forms found that the sampling worked as planned; an analysis using a chi-square test found that randomizing test sections also worked as planned. Next, we examined test content. An analysis using DIF found item stability. Several less stable test items were reviewed, but no consistent reason for the change could be found. ANOVA was used to examine test form equivalence. A significant difference was found between test forms, but effect size was miniscule. One test form was found to be problematic, supporting a previous decision to stop using it. Stability of scores over time was traced using a Shewhart control chart. Part of the overall variance was explained nicely by examinee demographics and other variables. When plotting the Shewhart control chart for the residuals, very few deviations over the years were found. Application of these new QC methods can facilitate fast and accurate identification of problems in past and current test forms. As an outcome of the study, automatic QC procedures were developed for use in continuously administered tests in general.

To what Extent are Local Norms Really Necessary for Every Spanish Speaking Country?

David Arribas-Aguila
TEA Ediciones, Madrid, Spain
e-mail: david.arribas@teaediciones.com

Over the last few years the use of psychometric tools is increasing gradually in some Spanish American countries. As a consequence, norms developed in Spain are the usual resource for interpreting the self-reports' scores in those countries, although Spanish American psychologists' demands for local norms are growing more and more. The main objective of this study is to analyze to what extent local norms requirement is justified by statistical differences between the self-report scores observed in both contexts. Raw scores from 5 well-known Spanish self-reports (CPS, CTC, compeTEA, CUIDA, and TPT) were analyzed in order to explore statistical differences by country. The samples were composed by all the administrations of these tests (N>180,000) made in the Spanish American countries, along with Spain, from 2012 to 2013. In most of the raw scores and traits analyzed, statistical differences between countries were significant, but small effect sizes were observed. Taking into account the differences observed, results suggest local norms development would be desirable for increasing the accuracy of the scores' interpretation. Nevertheless, the use of norms originally developed in

Spain does not lead to a significant bias for practical purposes. The appropriateness of global norms for all the Spanish speaking countries will also be discussed.

Comparing OECD PISA Reading in English to Other Languages

Mustafa Asil and Gavin Brown
The University of Auckland, Auckland, New Zealand
e-mail: m.asil@auckland.ac.nz

The OECD PISA results have been used as a catalyst for educational reform, even though the evidence for invariance across languages is weak. Scores obtained from tests that are subsequently adapted cannot be assumed comparable unless scalar equivalence across languages is present. The use of PISA across nations, cultures and languages has been criticized. The key criticisms point to the linguistic and cultural biases potentially underlying the design of reading comprehension tests and which raise doubts about the legitimacy of comparisons across language boundaries. Our research question was: What level of invariance is seen in the PISA Reading Comprehension test by language and culture relative to performance by the Australian English speaking reference group? To ensure equivalent sample sizes 500 students were randomly selected from PISA 2009 dataset for each country. MG-CFA based on MACS was used to examine levels of measurement invariance (MI). Changes in CFI ≤.01 were used to determine invariance. A scatterplot of ΔCFI against Δ□² (Chi-square) was used to identify patterns with respect to language and country. Measurement model had good fit for each country. Only Ireland, New Zealand, UK, English Canada and USA were invariant to Australia. Scatterplot with respect to language clarifies that language alone is not sufficient for invariance. For example, Trinidad and Tobago uses English but is not a high-wealth society. Spanish in Spain is relatively close to Australia compared to poorer countries of Latin America. Results indicate that complex factors to do with educational practice and socio-economic resourcing of education, rather than language per se do interfere with the MI of the PISA reading comprehension results. These results suggest that PISA reading comprehension tests are very much a product of a wealthy, Anglo-Commonwealth society and educational approaches to reading instruction.

International Differences in Personality: Smaller than Occupational Differences

Rob Bailey[1], Tatiana Gulko[1] and Marie Wendel[2]
[1]OPP Ltd, Oxford, United Kingdom; [2]Lund University, Oxford, Sweden
e-mail: rob.bailey@opp.eu.com

When a questionnaire is translated into another language, a number of factors make it difficult to judge whether it measures constructs equivalent to the original. Differences could arise from the translation process, cultural differences and different samples. These issues can be complicated further if the questionnaire includes questions designed to obscure their target from the questionnaire taker. This is the case with the 16PF personality questionnaire. The aims of this research were to: 1. Establish the equivalence of a range of different 16PF languages (e.g. UK, French, Dutch), using different English language versions as a benchmark (e.g. UK and US) 2. Create a single, international norm group for the 16PF 3. Compare differences in language/cultures with those found between different occupational groups. The study used multiple datasets (e.g. standardisation sets for UK, France, Netherlands and US), some of which included criterion data. Due to the various issues to examine, multiple statistical methods were used too: IRT-based Differential Item Functioning (DIF), Exploratory Factor Analysis and Confirmatory Factor Analysis, tests of difference at a scale level to examine differences in nationality/language, employment, and gender. The authors found surprisingly little DIF

between languages. Significant differences in mean scale scores were found between language versions, however these were small (generally little more than half a sten). The biggest significant differences in scale means were found between occupational groups (for example, people working in IT or health care). Lack of DIF suggests that the differences found between languages are not due to translation issues, but due to genuine cultural differences. These differences, however, were small when compared with two other yardsticks: 1. the Standard Error of Measurement, 2. occupational group differences. Our conclusion is that personality varies little between nations; however, cultural differences exaggerate the perceptions of personality differences.

Divergent Thinking as a Measure of Creative Potential: An Exploration with EPoC

Baptiste Barbot[1], Maud Besançon[2] and Todd Lubart[3]

[1]Pace University, New York, United States; [2]Université Paris Ouest Nanterre, Nanterre, France ; [3]Université Paris Descartes, Boulogne-Billancourt, France

e-mail: bbarbot@pace.edu

Divergent thinking (DT) tasks are the most commonly used measures of creative potential. However, their use as "pure" indicator of creative potential is controversial: Multivariate approaches have outlined the multidimensional hierarchic nature of creative potential (as opposed to unitary), which represents partly a generalized ability, partly a set of domain-specific abilities, and partly a set of task-specific abilities. We illustrate this phenomenon in a study extracting five variance components in multiple creative potential task scores: general variance (g), domain-specific, task-specific, thinking-process specific, and residual (i.e., measurement error). By doing so, we sought to estimate whether DT tasks scores are associated with enough g variance to legitimate their use as unique indicator of creative potential. 486 children and adolescents (mean age = 13.8, SD = 2.4) were administered the eight subtests of the Evaluation of Potential Creativity (EPoC; Lubart, Besançon & Barbot, 2011), measuring two key thinking-process clusters (divergent-exploratory, and convergent-integrative) in two domains (verbal and graphic). Test scores variance was partitioned in a structural equation modeling framework, followed by a Schmid-Leiman transformation. The contribution of each variance component on DT scores depends greatly on the task under consideration; some tasks being more "sensitive" to g, whereas others are more sensitive to domain-specific or thinking-process factors. As a general trend, results show that DT variance uniquely explained by g is rather limited. Most of the DT scores variance being process-specific (DT), with a large influence of domain-specific abilities for DT tasks in the verbal domain. This study suggests that DT tasks (especially verbal) may not be sufficient when used as a single indicator of "creative potential", a multidimensional construct in nature. It further outlines the need to measure creative potential with comprehensive test batteries sampling a range of creative tasks, domains and thinking processes.

The Underlying Structure of Academic and Cognitive Skills Used for The Diagnosis of Learning Disabilities

Anat Ben-Simon

National Institute for Testing & Evaluation, Jerusalem, Israel

e-mail: anat@nite.org.il

MATAL is a computer-based test battery for the diagnosis of learning disabilities (LD) in applicants to higher education institutions and in currently enrolled students. MATAL was developed as part of an endeavor to advance policy and procedure for standardizing and regulating the diagnosis of LD and the provision of test accommodations in higher education. The

MATAL Assessment Tools include: 20 tests (53 performance measures) which assess reading, writing, numeracy, attention, memory and visual perception. Of the 20 tests, 11 are used to diagnose language deficiencies, mostly those pertaining to reading and writing. Since 2007, over 14,000 persons have been diagnosed by means of MATAL. The objective of this study was to analyze the underlying structure of academic and cognitive skills used for the diagnosis of LD in general, and reading and writing disabilities in particular. Exploratory and hierarchal factor analyses were applied to 53 performance measures obtained by 2,652 persons who applied for LD diagnosed by means of the MATAL test-battery. The procedure was applied again to 31 performance measures of 11 language tests. The exploratory factor analysis applied to all MATAL performance scores indicated the existence of a principal factor that capitalizes on all the skills assessed by MATAL. This factor was interpreted as a general learning disability factor (LD g-factor) reflecting the efficiency of carrying out basic cognitive processes. Low to moderate correlations were observed between this factor and IQ scores. Exploratory, followed by hierarchal, factor analyses applied to the performance scores of the language tests categorized six language skills: lexical retrieval (naming), phonological decoding, text comprehension-accuracy, text comprehension-RT, orthographic memory and verbal fluency. The results of the study facilitate the understanding of the relations among a great variety of academic and cognitive skills used in the diagnosis of LD and the formation of integrated factor scores.

The Use of Data Visualization in Testing Reports

Brian Bontempo and Daniel Wilson

Mountain Measurement, Inc., Portland, United States

e-mail: brian@mountainmeasurement.com

Score reports are one of the most widely used and most important components of a test. One way to convey score information is through the skilled use of data visualization. Data visualizations are images such as tables and charts that convey information about a dataset. Recent innovations in data visualization suggest that this new technology may benefit the testing industry by making testing information more useable. The purpose of this study is to evaluate the use of data visualization in testing reports. Based on the data visualization theories of Tufte, Few, and Bontempo, the paper will identify ways in which data visualizations can be used to improve the quality of information contained in testing reports. These will include selecting the appropriate chart type and maximizing the Data/Ink Ratio (Tufte, 1983). The current use of data visualization in testing will be evaluated by analyzing a sample of examinee score reports and aggregate performance reports provided by the 282 certification testing programs in the US accredited by the National Commission for Certifying Agencies (NCCA). The quantity, type, and quality of the visualizations will be documented in addition to the type of certification program (health & wellness, business & finance, manufacturing/construction/trades) providing the reports. The quality of the data visualizations will be analyzed by estimating the Data/Ink Ratio and evaluating the appropriateness of the chart type (Y/N). Descriptive statistics will be calculated overall and by certification program type. We expect the results will indicate that data visualizations are included in about half of the reports provided by testing programs. We expect that, when included, the quality of the data visualizations is not optimal. In conclusion, this paper will provide guidance on the use of data visualization in testing reports and a good description of the current use of data visualization in testing.

A More Informative Group Percentile Rank

Brian Bontempo[1], Daniel Wilson[1], Philip Dickison[2] and Ada Woo[2]

[1]Mountain Measurement, Inc., Portland, United States; [2]The National Council of State Boards of Nursing, Chicago, United States

e-mail: brian@mountainmeasurement.com

The purpose of this study was to explore a group performance reporting metric (aka group score) that usefully combines the concepts of ranking, scaling, and norming. This metric, called Group Percentile Ranks (GPR), is obtained by calculating a measure of central tendency for each group and determining the percentile rank for the group by comparing each group's measure to the distribution of individual performance rather than the distribution of each group's measure. The shape of the distribution of group percentiles is approximately normal. This technique is currently used by the US nursing licensure examination to provide feedback to nursing schools on the performance of their graduates. The GPR method was evaluated from an empirical and theoretical perspective. Three years of data from the US nursing licensure exam were used to calculate traditional percentile ranks and group percentile ranks. The distributions were compared along the following criteria: ability to convey normative ordinal information, ability to convey normative interval information, impact of small program performance on other programs, quality of information at the extremes, and stability over time. Given rounding, the traditional method provided more ordinal differentiation amongst moderately performing programs while GPR provided more at the extremes. Although neither method was able to successfully convey equal interval information, the GPR method conveyed distance information better than the traditional method. The traditional method was negatively impacted by the performance of small groups while GPR method was not. The GPR method was more stable than the traditional method, a result attributable to the number of individuals being far larger than the number of groups. In conclusion, the GPR method provides a single metric which includes ranking, scaling, and norming information about a group's performance. As with the traditional method, caution should be exercised when making inferences for small groups.

Never Satisfied? Are the Most Motivated People the Hardest to Engage? Understanding Motivational Archetypes and their Implications for Organisational Commitment

Alan Bourne, Tony Li and Emma Stirling

Talent Q, Thame, United Kingdom;

e-mail: alanbourne@talentqgroup.com

Motivation is concerned with what drives people to behave in certain ways, in terms of their needs, aspirations and goals; how do people direct their energies and sustain effort? As with personality questionnaires, motivation questionnaires show clear differences in people's self-reported motivation across aspects of their work, be it acquiring wealth, personal development, having authority, etc., affording opportunities to use these tools to better engage people at work. Most research exploring motivation at work has taken a variable-centric approach, examining relationships between individual drives. Such an approach risks missing the point that different drives do not function in isolation from each other within a person. Taking a person-centric approach over a variable-centric can provide meaningful archetypes within multivariate data. Using latent class analysis, this research investigated whether there were coordinated configurations, or patterns of motivation in a working population sample which predicted significant differences in affective organisational commitment. The research revealed four clearly defined archetypes defined by key attributes such as emphasis on learning and pioneering/creating; need for authority, autonomy and acquiring wealth; motivation from

affiliation with others and job security; and a fourth group with lower overall levels of drive compared to the others. There were significant differences between these archetypes in terms of mean levels of organisational commitment, ranging from the group with the lower overall levels of personal motivation showing the highest commitment, through to a strongly learning-oriented archetype for whom commitment was the lowest. The results have significant implications for how organisations seek to understand and engage their people, demanding a need to avoid one-size-fits-all approaches and adapt to different psychological needs in a more flexible way.

The Added Value of Using Model-Based Classifications for Diagnostic Test Feedback

Laine Bradshaw

University of Georgia, Athens, United States

e-mail: laineb@uga.edu

Detection of students' strengths and weaknesses is an essential aspect of effective educational assessment systems (ETS, 2013). However, assessments have been criticized for not providing reliable, diagnostic feedback about what students know (Perie, Marion, & Gong, 2009).Often, subscores on content sub-domains aim to provide this type of diagnostic feedback, but fall short. Researchers caution the use of subscores (Tate, 2004) because they are often unreliable and lack added value over total scores (Sinharay, Haberman, & Puhan, 2007). Diagnostic classification models (DCMs) are newer psychometric models that hold promise for addressing diagnostic assessment needs (Rupp, Templin, & Henson, 2010); however, before implementing DCMs in practice to assign scores to students, evidence is needed to show the scores are reliable, comparable, and valid (AERA, APA, NCME, 1999). The purpose of this study is to (a) investigate the extent to which DCMs offer fine-grained feedback that is reliable and distinct and (b) contrast subscores and DCM classifications. Using a general DCM, we analyzed a US sample of 990 mathematics teachers' responses to the Diagnosing Teacher's Multiplicative Reasoning (DRL-903411) Fractions Test. This dataset is unique from others in the DCM literature because an interdisciplinary team prospectively constructed the Fractions Test for the explicit purpose of diagnosing teachers' mastery of four fine-grained attributes, or components of reasoning. In all cases examined, results indicted (a) DCM classifications offered an increase in reliability over subscores, and (b) inferences regarding attribute mastery based on subscores and DCMs estimates have a sizeable degree of inconsistency. The comparison of DCM estimates and subscores has not yet been examined empirically using data from a test designed to be diagnostic. Using the Fractions Test, we provided a practical illustration that DCMs provide one solution for efficiently providing multidimensional, diagnostic feedback that has acceptable reliability and added value.

Evaluating the Impact of an Intentional Item Pool Release

Chad Buckendahl[1], Russell Smith[2] and Jack Gerrow[3]

[1]Alpine Testing Solutions, Las Vegas, United States; [2]Alpine Testing Solutions, Orem, United States; [3]National Dental Examining Board of Canada, Ottawa, Canada

e-mail: drcbuck@gmail.com

Security considerations for testing programs continue to influence policy and practice. Programs commit efforts to security considerations based on prioritizing prevention, detection, and enforcement activities. Prevention generally offers testing programs the greatest opportunity to exert control over their materials. However, these good intentions may have an unintended effect of causing individuals to focus even more on trying to determine test content. In such instances, the benefit of creating a large item pool may be counteracted by the limited number of

items that appear on an operational form. An alternative approach to maintaining control of an item pool has been described as "item pool flooding" (ATP, 2013). Multiple variations of the concept have been postulated with one goal of trying to exhaust candidates who simply try to memorize all the items in the pool. This paper describes a study of an empirical evaluation of the stability of item and test form characteristics that was conducted as a result of a policy decision by a licensure examination program to release an approximately 7,000 item pool in 2009. Results are based on approximately 3,750 candidates who took the exam from 2007-2012. Analyses focused on test- and item-level performance and item level change across years. Of the 126 items that were reused on operational forms following the release, only one exhibited a statistically significant difference in difficulty. This study is a first step in evaluating innovative strategies in test security. Specifically, methods for prevention and detection of item exposure have relied predominantly on maintaining control over information. Differential resources for implementing these activities can yield unintended consequences. Practitioners are cautioned against making a leap to release their item pools without further exploration of the topic in the context of their testing program. Recommendations for considering this strategy given programmatic characteristics are also provided.

Common Practices in the Adaptation of Psychological Tests

Yesim Capa Aydin

Middle East Technical University, Ankara, Turkey

e-mail: capa@metu.edu.tr

The number of psychological tests being adapted into various languages has increased immensely during the past two decades. However, test adaptation is not often carried out properly, even limited to "text translation" in some instances. Acknowledging the need for technical literature in the field, the International Test Commission (ITC) prepared and formally presented guidelines for test translation and adaptation first in 1999 at a conference in Washington, DC. Gregoire and Hambleton (2009) consider guidelines as "rules of good practice" rather than "list of psychometric procedures" (p. 76). There are 22 guidelines under four sections. Rationale for each guideline is included along with the steps to meet guidelines. Considering the fact that approximately fifteen years have passed after the first publication of Guidelines, the present study will provide a review of methodological issues involved in the test adaptation studies. In order to identify common practices, I will review two well-known psychological journals published between 2000 and 2013. The following issues will be examined: (1) the competency (cultural/professional/linguistic) of the selected translators; (2) the number of translators; (3) translation design (backward and/or forward translation); (4) the extent to which ITC guidelines is used and cited; (5) data collection design (monolingual and/or bilingual design); (6) the number of data collections; (7) the adequacy of sample size in each data collection; (8) statistical analyses for psychometric properties. Preliminary review of studies indicated that researchers commonly used backward translation with the number of translators ranging from 1 to 5. The linguistic competency of translators is emphasized the most, while cultural competency is addressed in a small number of studies. Only a few number of studies refer to the ITC guidelines. Monolingual designs were preferred more than bilingual designs. Factor analysis, Cronbach's alpha, and correlational analysis were the most frequently used statistical approaches.

Documenting Impact of Reliability on Estimates of Classification Accuracy and Consistency of the CELPIP-G Speaking and Writing Scores

Michelle Y. Chen[1] and Amery D. Wu[2]

[1]The University of British Columbia / Paragon Testing Enterprises, Vancouver, Canada; [2]The University of British Columbia, Vancouver, Canada

e-mail: michellec2004@gmail.com

Classification accuracy (CA) and consistency (CC) are major concerns in criteria-referenced score use, such as the Canadian English Language Proficiency Index Program® General (CELPIP-G) Test. CELPIP-G test takers are classified to 12 levels of Canadian Language Benchmarks (CLB) for immigration/citizenship purposes. Although several methods are available for CA/CC, none were designed specifically for performance assessment such as the CELPIP-G speaking and writing assessments. Livingston and Lewis' (LL, 1995) method claimed that it can be applied to any scoring system based on score reliability. However, the LL method was found to be sensitive to the type of reliability estimates chosen (Deng, 2011; Wan, Brennan, & Lee, 2007). The study's main purposes were: (1) to estimate CA/CC for the CELPIP-G speaking and writing assessments through the LL method and (2) to discuss how the choice of different reliability coefficients for LL might impact the CA/CC estimates for the CLB cuts. There were 8 and 2 tasks for the speaking and writing assessments, respectively. Each task was rated on four dimensions (1- 5) by two independent raters. Two methods were used to estimate reliability: phi coefficients of generalizibility theory for absolute decision based on task scores and Cronbach's alpha internal consistency among the dimension scores (64 and 16 scores for speaking and writing, respectively). The program BB-CLASS (Brennan, 2004) was used. Very good CA/CC rates were obtained for all cuts ranging from 0.85 to 0.99. Higher CA/CC was associated with higher reliability. For speaking, CA/CC based on alpha yielded higher values than those based on phi because alpha estimated higher reliability than the phi. The reverse pattern was found for writing because phi estimated higher reliability than alpha. The findings confirmed that the LL method was sensitive to the choice of reliability, although the effect was small (max difference=0.05).

The Response Process of Applicants' Faking on Personality Test: Using Mixed-Method to Explore the Cognitive Processing Mechanism*

Jiyue Chen and Jianping Xu

[1]School of Psychology, Beijing Normal University, Beijing, China;

e-mail: 672077809@qq.com

Employers fear that applicants will fake if they take a personality test in personnel selection condition, which has inspired much research. But contradictory results provided little evidence about fakers' underlying cognitive process. A better understanding of their cognitive process can help employers to find more specific and effective ways to control faking. To explore applicants' cognitive process of faking on personality test in personnel selection condition. Mixed-method design was adopted to study the response process of faking. Study 1 was a within-subject experiment aimed at the cognitive process of faking. The BFI-44 was administered to 200 college student applicants in both honest condition and selection condition. Applicants' item response, reaction time and eye-movement indicators were recorded by Tobbi 120. Study 2 focused on the thinking process of faking. Twenty subjects participated in an interview that was guided by grounded theory. Interview content consisted of the subjects' motivation and cognition when they were faking. Faking is easier than honest answering on the items with job desirability. Applicants reacted faster, had more fixations on the 2 extreme response options, and they fixated on these more directly after having read the question. Faking is more difficult than honest

answering on the items with job undesirability. Applicants reacted more slowly and had more fixations on moderate options. They didn't fake on items unrelated to job desirability in selection condition. Contents analysis showed that some applicants would fake on some items strategically based on job desirability to increase their employment probability. But some applicants would not fake. Applicants fake on the personality test with the adopted schema model based on job desirability instead of another three mutually contradicting cognitive processing models. The present study further explored the cognitive process of faking and gained some insights by way of mixed-method.

In Search of Culture Fair IQ Tests: A Comparison between South African and British Students on The WAIS-III

Kate Cockcroft[1], Tracy Alloway[2], Evan Copello[2] and Robyn Milligan[3]

[1]University of the Witwatersrand, Johannesburg, South Africa; [2] University of North Florida, Jacksonville, Florida, United States; [3]Johannesburg, South Africa

e-mail: kate.cockcroft@wits.ac.za

There is ongoing debate regarding cross-cultural influences on IQ tests. This paper reports on the use of the Wechsler Adult Intelligence Scale-III (WAIS-III) with an English second language, multilingual, low socio-economic group of young South African adults. The objective was to compare their performance on the WAIS-III to that of a British, monolingual, higher socio-economic group of young adults. Both groups were assessed under similar one-on-one conditions, by a psychologist, on the university campus. A series of MANOVAs revealed that the British group significantly outperformed the South African group on the knowledge-based verbal (as well as some nonverbal) subtests, while the South African group showed particular strengths in terms of processing speed. A discriminant function analysis confirmed that higher Verbal Comprehension scores characterized the British students, while higher Processing Speed scores typified the South African students. The latter finding may be an effect of the South African students' longstanding multilingualism, which requires constant switching between, and monitoring of, several languages. The groups showed no significant differences on the working memory subtests of the WAIS-III, which appear to be the least culturally biased. This suggests that tests of working memory offer great promise in the search for fairer assessment practices.

What Cognitive Levels do Students Really Use when Solving the Items? Cognitive Interviews As an Efficient Tool in Test Validation

Natalija Curkovic

National Centre for External Evaluation of Education, Zagreb, Croatia

e-mail: natalija.curkovic@ncvvo.hr

When constructing knowledge tests, cognitive level is usually one of the dimensions comprising the test specifications with each item assigned to measure a particular level. There are many concerns in current literature about existence of predefined cognitive levels. Since the unidimensionality is one of the requirements when constructing the knowledge tests, standard statistical procedures like factor analysis and structural equation modeling usually cannot show which item measures which cognitive level. The aim of this paper is to to investigate: (I) can statistical methods confirm presence of different cognitive levels?; and (II) what cognitive levels do students really use when solving the items? For the purpose of the research, a Croatian final high-school Mathematics exam was used (N = 9626). Confirmatory factor analysis and structural regression modeling were used to test three different models. Additionally, 16 senior high-school

students individually solved the Math test and after each solved item, they were asked to explain the cognitive processes used to reach their solution. The whole process was audio-recorded and analyzed afterwords. Structural equation modeling techniques did not support existence of different cognitive levels in this case. The analysis of students' cognitive interviews involved 582 students' answers. After classification into categories of cognitive levels, answers were summed and compared to the test blueprint. The agreement rate with the test developers' cognitive level classifications was 62% (28 of from 45 items). Agreement rate between the test makers judgments of cognitive levels measured by the items and students answers indicates the existence of recognizable cognitive levels. Cognitive interviews turned out to be an efficient tool in the test validation procedure which cannot be replaced by the standard quantitative methods.

Toward Maximizing the Likelihood of Comparability and Equivalence: A Framework for Adapting the Psycholexical Methodology

Lina Daouk-Oyry[1], Pia Zeinoun[2], Fons van de Vijver[2] And Lina Choueiri[1]

[1]American University of Beirut, Beirut, Lebanon; [2]Tilburg University, Tilburg, Netherlands

e-mail: linadaouk@gmail.com

The first big shift in cross-cultural assessment towards better practices was from translation to adaptation of tests in order to increase the likelihood of achieving equivalence between their multi-lingual versions. The conversation in the field of personality assessment, however, has moved beyond questioning the equivalence between the tools of assessment, rather, the model upon which they were constructed. Researchers in different countries have re-investigated the factor structure of personality based on the analysis of their respective lexicons and found differences with the original English language-based Five Factor Model. However, the methodologies that these researchers are utilizing may hinder future comparability between their findings, and consequently hinder or limit our understanding (and measurement) of personality structure around the world. The aim of the study was to propose a framework for standardizing the procedure of the psycholexical approach in order to maximize the likelihood of method equivalence. We develop our framework using the Arabic psycholexical study as a case in point. We rely on the challenges faced while conducting this study considering the available literature to support the application of this method. The resultant framework presents protocols for adapting the psycholexical methodology onto a specific language to maintain a high potential for comparability with studies in different languages in the future. The mixed emic/etic approaches commonly utilized in cross-cultural assessment bring with them many language specific challenges that may limit future generalizability or the attainment of global perspectives in the field. This model provides researchers the opportunity to conduct local research while making sure they keep the global comparability in mind.

Global and Local Challenges in English Proficiency Test Scores Use in Graduate Student Admissions

Slobodanka Dimova[1], April Ginther[2] and Catherine Elder[3]

[1]University of Copenhangen, Copenhagen, Denmark; [2]Purdue University, West Lafayette, IN, United States; [3]University of Melbourne, Parkville VIC, Australia

e-mail: plq379@hum.ku.dk

Current theoretical discussions on consequential aspects of test validity have led to increased interest in how test scores are used and understood by stakeholders within particular domains. Adopting an instrumental case study approach, this study examines levels of knowledge about the English language tests (TOEFL, IELTS, and PTE) used for selection in two academic contexts

and the uses of test scores in local decision-making by graduate faculty. Data for the study were gathered via an online survey and follow-up interviews probing the basis for participants' beliefs, understandings, and practices. The presentation focuses the results of the 50-item survey completed by respondents from Research 1 universities across three continents: 232 respondents from U.S., 246 respondents from Australia, 240 respondents from Europe, and 45 follow-up interviews at all institutions. At a global level, responses reveal that while admissions practices vary considerably across disciplines, English test scores, once entry-level requirements are met, tend to have very limited impact on admissions decisions as compared to evaluations based on other kinds of evidence, including telephone interviews, writing samples, and recommendations. In all three contexts, respondents emphasized (1) the importance of English for academic success; (2) dissatisfaction with current levels of English among graduate students; (3) limited knowledge about or understanding of the major English tests used for selection; and (4) a concomitant lack of preference for any one of the 3 tests accepted for admissions purposes. Despite limited use of language proficiency test scores by decision makers, there is also evidence that users are quite savvy about what matters for academic success. discuss the global practical challenges associated with ensuring that critical information about language tests is locally tailored to fit the needs of different audiences as well as the complex mechanisms and distributed responsibilities involved in the student selection process.

Compare the Main Teacher Competency of Secondary Schools in China and Catalonia

Shujing Ding

Universitat Autònoma de Barcelona, Barcelona, Spain

e-mail: yangchun.xin1986@gmail.com

This study used questionnaires to survey 808 in-service secondary school teachers in China and Catalonia. 408 of them are Chinese teachers and 400 of them are Catalonia teachers. We have also interviewed 46 teachers and 16 managers of secondary school. There are 24 Chinese teacher and 22 Catalonia teachers, 8 Chinese secondary school managers and 8 Catalonia secondary school managers. Its findings are as follows: (1) we have developped two questionnaires for testing secondary school teachers in China and Catalonia. (2) The Chinese secondary school teachers think that the most important teacher competencies is how to help students to learn knowledge, which is different from Catalonia secondary teachers. (3) In China, secondary school managers think the most important five teacher competencies are teaching skills, ability to learn specialized knowledge, ability to grasp secondary teaching materials and methods, communication. (4) In Catalonia, secondary school managers think Personal and Professional Development, Values-Professional , Knowing the student, Learning and Teaching Process, Monitoring and Evaluation are the most important five teacher competencies.

Differential Performance on Technology Enhanced Items

Tonya Eberhart[1] and Marianne Perie[2]

[1]University of Kansas, Lawrence, KS, United States; [2]CETE at the University of Kansas, Lawrence, KS, United States

e-mail: teberhart@ku.edu

Currently in the United States with the push towards college- and career-ready standards, test developers are experimenting with new item types in K-12 testing. Specifically, technology-enhanced items allow students to interact with the item, dragging and dropping text or objects, reordering images, graphing number lines, providing visual representations of fractions, and

highlighting text within passages. These item types are presumed to provide more authentic interaction with the construct being assessed, allowing for better and more efficient testing. However, there have been few studies of these items with the K–12 population. Moreover, these items are given on a multitude of devices (tablets, laptops, and desktops) using a multitude of platforms (Google chrome, ipad, Mac, and IBM based). And students interacting with the items have a multitude of backgrounds and experience with the devices. One particular consideration with technology-enhanced items is the difference between using a mouse to manipulate an item on a desktop/laptop and using a finger on a tablet. It is important to examine whether this manual difference results in any difference in performance, and whether any differences in performance are specific to students with particular characteristics. This paper will explore the interaction between the student, the device, and the item to determine if there are any differential item effects that need to be mitigated.

Cross-Country Differences in Reported Test-Taking Effort: A Measurement Invariance Study

Hanna Eklöf
Umea University, Umeå, Sweden
e-mail: hanna.eklof@edusci.umu.se

When groups are to be compared on latent variables, the issue of measurement invariance is important to consider, i.e., we need to know whether a measure seems to assess "the same construct in the same way" across groups. This is true also in the context of international comparative studies like TIMSS and PISA, and the self-report scales that are used in these studies. The purpose of the present study was to exemplify and discuss this issue by evaluating the measurement invariance of an effort scale used in Sweden and Russia in TIMSS Advanced 2008. TIMSS is a low-stakes test, causing concern that students may not be motivated to do their best and that this might be a threat to the validity of inferences made from test scores. To account for this possible source of construct-irrelevant variance, we used a six-item self-report scale assuming to measure reported effort and motivation to do one's best on the test. Results indicated large differences between Sweden and Russia in terms of reported effort and motivation. To investigate whether these differences seemed to be "true" differences in reported effort and motivation, a multigroup confirmatory factor analysis (MGCFA) was performed. According to the MGCFA, the scale was reasonably invariant on a configural (factor structure) and metric (factor loadings) level, but not on a scalar (intercept) level, indicating that there may be differences (bias) in how students respond to the items in the effort scale in Sweden and Russia, respectively. Results suggest that we need to acknowledge that self-reports not necessarily only reflect "true differences" in the latent variable but may also reflect different response styles across countries, but also that full strong invariance might be difficult to reach when we sample subjective perceptions across cultures.

Invariant Measurement with Raters

George Engelhard
The University of Georgia, Atlanta, GA, United States
e-mail: gengelh@uga.edu

Raters appear in applied settings that range from high-stakes performance assessments in education through personnel evaluations in a variety of occupations to functional assessments in medical research. This study utilizes principles of invariant measurement (Engelhard, 2013) combined with lens models from cognitive psychology (Cooksey, 1996) to examine quality of judgmental processes that arise in rater-mediated assessments. Specific research questions

are: 1. Are the raters interpreting the domains as intended? 2. Are the raters interpreting the category structure as intended? 3. Are student scores invariant over raters (order and rating consistency)? Data from a large-scale writing assessment (N=6,360) are used to illustrate guiding principles for assessment systems based on human judgments. Rasch measurement models are used in the development of psychometrically sound rater-mediated assessments based on the principles of invariant measurement. The preliminary analyses suggest that some raters differ in their interpretation of four writing domains (Ideas, Organization, Style, and Conventions). The data also suggest an interaction effect between domains and the use of the categories by the raters. There also appears to be significant variation in the consistency of student scores across raters. The principles of invariant measurement can be used to identify sources of construct irrelevant variance in rater-mediated assessments. Quality control procedures can minimize these rater errors. Future research on quality control procedures is a promising area for improving the psychometric quality (reliability, validity and fairness) of ratings assigned by raters. References Engelhard, G. (2013). Invariant measurement: Using Rasch models in the social, behavioral, and health sciences. New York: Routledge. Cooksey, R. W. (1996). Judgment analysis: Theory, methods and applications. San Diego, CA: Academic Press. Linacre, J. M. (2007). A user's guide to FACETS: Rasch-model computer programs. Retrieved from www.winsteps.com.

Cross-Cultural Differences in Ambivalent Response Style

Yehuda Esformes[1], Sompong Virakananon[2] and Alex Rodan[3]

[1]Amelia Systems, Tel Aviv, Israel; [2]Insight Advice, Bangkok, Thailand; [3]Til QUEST recursos humanos, Barcelona, Spain

e-mail: yehudaes@gmail.com

There is a large body of research in cross-cultural psychology that describes notable differences in response styles between East Asian and western cultures. Most studies focus on a number of well-known response styles such as acquiescent responding, extreme responding, midpoint responding and socially desirable responding. Ambivalent responding as a response style has been much less investigated. Some exceptions are the work of Spencer-Rodgers et al. (2004) and Hamamura et al. (2007) who found that, compared to western participants, East Asian participants are more ambivalent in their response choices. These differences have been attributed to cultural differences in dialectical thinking. The purpose of this study was to investigate differences in ambivalent response style between Thailand, a large community that is rarely addressed in psychological research, and Spain. The data was drawn from two datasets of 287 Spanish and 3106 Thai participants that responded to a multi-trait self report inventory for the purpose of a vocational counselling. Ambivalence was operationalized in two different ways: 1. Contradictory responses given to pairs of highly correlated items. 2. The frequency of endorsing the relatively ambiguous response category "Not so true" compared to the other more clear and unambiguous choices ("Definitely true", "True" and "Definitely not true"). ANCOVA on the ambivalent response style measures, controlling for age, educational level and personality trait-level, revealed no gender differences, but significant cultural differences in almost all measures of ambivalent responding. Thai respondents made more contradictory choices and selected more frequently the "Not so true" category than their Spanish counterparts. The various patterns of differences observed in our study are explained on the basis of Spencer-Rodgers et al. approach, viewing the roots of the differences as stemming from religious and philosophical traditions.

Creation of Russian Version of Psychodiagnostic Technique “LOT-R”*

Kseniya Evnina and Diana Tsiring
Chelyabinsk State University, Chelyabinsk, Russia
e-mail: evnina@mail.ru

The object of our research was adaptation of psychodiagnostic technique "LOT-R" to Russian population. This research was done with the permission of the authors of the original method "Life orientation test" (M. Scheier and C. Carver). The first stage of adaptation was the Russian version of the questionnaire creation. As M. Scheier and C. Carver pointed out, optimists and pessimists differ in the ways that have a big impact on their lives. We have studied the characteristics of the Russian mentality in this issue. The peculiarities of "Russian optimism and pessimism": the belief in miracles, aspirations for the future as opposed to the immediacy and optimistic attitude, not associated with material well-being have been taken into account when formulating new questions. The study established measure of internal consistency of the method (0.926), correlation between the original and additional parts of the questionnaire (0.807) and reliability of the half-splitting (0.907) at a high level of validity. An important step was the establishing of retest reliability coefficient for which the adapted technique is at a high level of confidence (0.839). In our work theoretically obvious and construct validity procedures have been described, and studied with respect to the validity of the competitive method of M. Seligman «ASQ» (0,516). The final stage includes standardization and setting norms of the test results for Russian population. An adapted version of the method allows organizing multicultural research, covering the phenomena of optimism and pessimism.

Application of the Argument-Based Approach to Validity Evaluation

Ellen Forte
edCount, LLC, Washington, United States
e-mail: eforte@edcount.com

Tests must be purposely designed to yield scores for particular uses and test scores must be accompanied by some evidence that supports those intended uses. This paper addresses current expectations as defined in the Standards for Educational and Psychological Testing (AERA/APA/NCME, 1999) in the context of the argument-based approach to validity evaluation (Kane 2002, 2006, 2013) and as manifested in real-world settings. This paper articulates means by which test developers and users can establish a comprehensive body of evidence, based on professional standards, using modern validity theory to guide the evidence collection and evaluation process. Further, this paper illustrates how has been achieved in two settings in ways that support the needs of US states to submit evidence for review by the US Department of Education. This paper presents and explains the rational for a validity evaluation framework that test developers and users can apply to student achievement tests or adapt for other applications. In addition, this paper includes components from two sets of real-world technical documentation. As this paper does not reflect a research study, but a model for how to identify necessary studies to address key validity questions and to interpret, organize, and evaluate the results of those studies, there are no results, per se. The manner in which test developers and users gather, evaluate, and describe evidence of test quality typically does not result in a coherent body of evidence that relates clearly to the argument-based approach or any other logic framework that supports validity judgments. The framework offered in this paper addresses these issues and provides a means for both planning and evaluating the validation process.

Adapting the d2-R from Paper and Pencil to Online Administration

Bastian Funken[1] and Herbert A. Meyer[2]

[1]Hogrefe Verlag Gmbh & Co. KG, Goettingen, Germany; [2]artop - Associated Institute of the Humboldt University of Berlin, Berlin, Germany

e-mail: bastian.funken@hogrefe.de

The Test d2 - Revision (d2-R) is a further development of the established d2 Test of Attention. The paper-and-pencil form of the d2-R was published in 2010. The online adaption of the d2-R was recently published within the Hogrefe TestSystem 5, a modern test suite for online assessments. To preserve both the ease of use as well as the psychometrical strength of the d2-R the iterative design process to develop the digital version of d2-R was guided by a close orientation to the paper-and-pencil form. We conducted usability testing with eye tracking to gather feedback from test users to check design decisions. The most difficult challenge was the absence of a diagnostician in the online setting. Therefore our main were the evaluation of the instructional quality and the correct usage of the test administration (line-by-line working). One finding was that a reliable testing requires a training phase more than a narrative instruction. The narrative instruction could be reduced to a minimum. Another finding was that all test users consistently worked line by line if the arrangement of items was determined by the law of proximity. While the item difficulty generates a strong variance between regular users, very experienced users were able to solve the task with high precision over time. Deeper analysis of eye movements showed details of the faster information processing which led to the better performance. The results show that the online adaptation of the d2-R works as reliable and valid as the paper-and-pencil version and above that, can be used in distributed settings without a diagnostician. In addition the specific implementation of the d2-R allows for a usage with modern browser applications and tablet computers. Highlight videos and results of the current standardization study will be part of the presentation.

Reverse Scoring Items Effect on Psychometric Properties of Tests

Ángel García-Pérez, Gema Alonso, Ignacio Pedrosa and Eduardo García-Cueto

Universidad de Oviedo, Oviedo, Spain

e-mail: rydenone@gmail.com

Personality and attitudes tests are used in many contexts of psychological assessment. Often, a common strategy in these types of tests is to include both items that agree with its content involving ranked in the variable measured, such as items in which disagreement with them also carries ranked in the same variable. In fact, many tests builders advise to do it to avoid acquiescence, that is, to reduce the tendency of the subject to show its agreement with the item content. However, having worded items in both "directions" can affect the test score of the participant as well as the psychometric properties of the test. The target of this work is to study the influence of the reversed items on the psychometric properties of the scale used in terms of reliability and validity. Three forms of self-efficacy test were developed, A, B and C: "A" test items were written in affirmative sentences, "B" the test items were worded as reversed sentences, and "C" test items were worded in both. In order to eliminate the memory effects, between all the three tests administration, a period of time long enough was allowed. The effect of the application order was controlled. The results showed the reverse scoring affects the psychometric properties of the test, both from the classical perspective and from IRT models. Items redirection affects several aspects such as test reliability and validity, and items indexes such as discrimination, difficulty and homogeneity.

Methods for Checking Fit between Angoff Ratings and IRT Response Data

Adam Gesicki[1], Amery Wu[2], Jake Stone[3] and Michelle Chen[2]
[1]UBC / Paragon Testing Enterprises, Vancouver, BC, Canada; [2]UBC, Vancouver, BC, Canada
[3]Paragon Testing Enterprises, Vancouver, BC, Canada
e-mail: gesicki@paragontesting.ca

The Angoff procedure suggests a cut score on the observed scale, usually requiring a transformation of Angoff cut scores to IRT true scores for the purpose of reporting of individuals' performance. Most transformation methods suggested to-date follow the common presumption that the Angoff Pannelists' ratings fit the chosen IRT model (goodness-of-fit assumption). However an Angoff (1971) cut, due to rater subjectivity, requires evidence for its credibility. Having good fit points to the trustworthiness of the Angoff cut. This study has two purposes: (1) to examine the goodness-of-fit assumption using the test suggested by Kane (1987) with real data and, (2) to demonstrate a new method for checking the goodness-of-fit assumption using a differential item/test functioning (DIF/DTF) technique. Kane's (1987) statistic follows chi-square distributions with df=n-1 for the goodness-of-fit assumption (p. 336). Six Angoff cuts were obtained each from 20 and 22 Pannelists for the listening and reading pilot tests of the revised CELPIP-G: a requisite test of language ability for immigration to Canada. The 2PL IRT model was chosen to calibrate 500 test-takers' responses. The new method is suggested based on the rationale that, if the goodness-of-fit-assumption holds, an item/test would function the same between the two groups (test-takers and Pannelists), given a minimally competent test-takers' true abilities. This rationale can be tested using the typical IRT DIF/DTF method. At the test level, the result of DTF will be compared to Kane's method. DIF significance and effect size will be reported; hence, test-level (mis)fit can then be examined through item (mis)fit. Advances are gained through the new method's capacity to concurrently model both data sets and directly checking the goodness-of-fit assumptions by testing DIF/DTF simultaneously.

Selecting Civil Servants across 28 European Countries: Psychometric Challenges and Solutions

Gilles Guillard
European Personnel Selection Office, Brussels, Belgium
e-mail: Sandy.lecluse@ec.europa.eu

The European Personnel Selection Office (EPSO) delivers a staff selection service on behalf of the Institutions of the European Union. For each selection process candidates from the 28 member states are assessed in order to select the best for possible recruitment as EU officials within the Institutions. Around 40 000 graduates apply every year for less than 300 posts of generalists and undergo a two stages selection process: a pre-selection phase with CBT and an assessment centre. The pre-selection phase consists of tests assessing verbal, numerical and abstract reasoning as well as situational judgement. Verbal and numerical reasoning is delivered in all 24 official languages of the European Union. The major psychometric challenges for CBT are translating and calibrating the items and assembling them into equivalent test forms. In order to best meet these specific challenges of multilingual large-scale testing EPSO has developed innovative and sophisticated in-house solutions involving comprehensive item analyses based on the Rasch model, equating across the 24 languages and a test form assembling algorithm. By using these solutions EPSO can ensure fair and equal treatment of all candidates.

Comparison of Software Packages for Uni- and Multi-Dimensional IRT Model Estimation

K. Chris Han[1] and Insu Paek[2]

[1]Graduate Management Admission Council, Reston, United States; [2]Florida State University, Tallahassee, Florida, United States

e-mail: khan@gmac.com

The concept of multidimensional IRT (MIRT)—the mathematical equivalent to existing factor analysis approaches—has circulated in the field for decades, but had yet to reach critical mass in terms of relevant research studies and application development. The last five years, however, have witnessed great progress in the development of several new software tools for MIRT modeling designed to take advantage of the latest available computing technology. Because these MIRT software tools differ significantly in terms of their performance and user experience, researchers and practitioners could benefit from a fair, comprehensive comparison of these tools to make more informed decisions in choosing the right MIRT software tool for their specific needs. This study presents an evaluation and comparison of the most recent commercially available MIRT software packages and their estimation algorithms. These packages include: IRTPRO 2.1 (Cai, Thissen, & du Toit, 2011), mPlus 7.11 (Muthén, & Muthén, 1998–2012), FlexMIRT (Cai, 2012), and EQSIRT (Wu & Bentler, 2013). The study examines the performance of model recovery via a series of simulations based on three approaches for latent structuring— within-item MIRT, between-item MIRT, and a bifactor model. Unidimensional IRT cases were studied as well. The simulation studies focused on realistic conditions and models that researchers and practitioners are likely to encounter in practice. The results showed that the studied software packages recovered the item parameters reasonably well but differed greatly in terms of the types of data they could handle and also in time required for estimation completion. This study makes no attempt to identify any particular solution as the "best" among the studied software packages. Instead, it offers suggestions for which model would be more appropriate and work more effectively under specific situations based on the varying strengths and shortcomings of each MIRT software package.

Life may not be Linear, so what about Selection? Evidence for Nonlinear Relationships between Personality and Performance

John Hackston[1] and Swati Kanoi[2]

[1]OPP Ltd, Oxford, United Kingdom; [2]University of Oxford, Oxford, United Kingdom

e-mail: john.hackston@opp.com

Practitioners often assume nonlinear relationships between personality and job performance; this underlies many recruitment techniques, from 'ideal profiles' to danger-zones, qualification grids, and profile matching (Kulas, 2013). Yet research has frequently failed to find evidence of curvilinear relationships between job performance and personality or has found only very slight effects (Whetzel et al, 2010). This study investigates the extent to which nonlinear models offer advantages over linear, using the relationship of personality scales to performance ratings. Much existing literature has been theory based, and we attempt to close the practice-theory gap. 279 managers attending leadership development took part, rating themselves, and asking others to rate them, on job performance using a 360° instrument (Benchmarks; CCL, 2001). They also completed the 16PF questionnaire (Russell and Karol, 2002). 16PF dimensions were compared with self- and boss ratings of leadership, derailment, promotability, and performance. Bivariate correlations were computed (using 16PF primary factors rather than the FFM, as practitioners will typically use the former in real-life selection). Regressions established which model (linear, quadratic, cubic, power, S, logistic or exponential) best fitted each relationship. Where a model

demonstrated a R2 change which was less than 0.01 over a simpler model, the simpler was preferred (Whetzel et al, 2010). The results provide some support for nonlinear effects. Nonlinear models mostly offer only modest returns over linear, but in some cases a nonlinear relationship offers significantly more, especially where little linear relationship exists. Where a strong linear relationship is present, nonlinear models may give a slightly better theoretical fit to the data, but add little incremental value and involve unnecessary complications in practice. The results show that practitioner approaches can be simplistic, but that non-linear analysis can be applied to real-life selection processes.

Culture or Personality? Individual Differences in Cultural Orientation across Countries

John Hackston and Gaby Walker
OPP Ltd, Oxford, United Kingdom
e-mail: john.hackston@opp.com

Previous research has shown that culture varies across countries, and many researchers have developed culture models, notably Hofstede (1980) and Trompenaars (Trompenaars & Hampden-Turner,1998). The Cultural Orientations Framework (COF; Rosinski, 2003) takes aspects of Hofstede, Trompenaar and others to derive a set of cultural orientations - an inclination to think, feel or act in culturally determined ways. This study sets out to investigate the structure of culture, as measured by the COF, and establish the extent to which individual differences in cultural orientation are the same as or different from individual differences in personality. Data was collected online between April and December 2013, from 1,056 participants who already knew their personality type (as measured by the Myers-Briggs Type Indicator® questionnaire (Myers et al, 1998). Cultural orientation was measured using the Cultural Orientations Framework (Rosinski, 2003); this measures both orientation and perceived ability across 17 dimensions. Although used in coaching, little empirical research has previously used the framework. The results showed a generally high degree of agreement between cultural orientation and perceived ability on most dimensions, with some exceptions (for example, use of scarce resources). Factor analysis of both orientation and ability showed structures more reflective of personality models than of cultural models, and these factors showed clear relationships to personality dimensions. However, differences in orientation between respondents resident in and originating from different countries accounted for significant variance over and above that accounted for by personality scales. The results show that both cultural and personality factors affect our orientation and behaviour. We will illustrate the practical implications of this and explore techniques for working with people from different cultures.

US Assessment Consortia: Status and Progress in Assessing Deeper Learning

Joan Herman
UCLA/CRESST, Los Angeles, CA, United States
e-mail: herman@cse.ucla.edu

The United States invested $400 million in two state-consortia to develop assessment systems aligned with Common Core State Standards for English Language Arts (ELA) and Mathematics. The new systems are intended to drive K-12 school reform to support students' deeper learning -- the thinking, problem solving and communication skills students will need for college and career readiness. This paper summarizes the progress of both consortia in achieving their charge, concentrating on the extent to which the tests will address deeper learning goals. Consortia plans for assessing deeper learning are also compared to current prominent national and international

exams. Study methodology used Webb's four-point Depth of Knowledge (DOK) Framework (Webb et. al. 2005) to analyze representation of deeper learning in consortia's content, item and task specifications, sample items and test blueprints. Consortia results are compared with DOK analyses of current state tests in the US, PISA, TIMMS, PIRLS, IB, and US NAEP and AP exams (Yuan & Le 2012, in press). Analyses reveal that 1/3 or more of both consortia's tests will reflect DOK levels 3 or 4, in contrast to current state tests and respected national and international exams. Across all tests, DOK is higher in ELA than in mathematics. Results indicate that consortia tests will be a substantial step ahead in assessing the kinds of deeper learning needed for 21st century success. Time will tell the extent to which the consortia will be able to stick to these rigorous targets: field testing and operational test forms are yet to come. Study findings also suggest a need for better frameworks for analyzing cognitive complexity that simultaneously also could serve item and task design purposes.

Manifest Scores in Practical Applications, Latent Scores in Research: Does it matter?*

Anne Herrmann

Kalaidos Fachhochschule Schweiz, Zürich, Switzerland

e-mail: inbox.anne@gmail.com

In the past, most psychometric tools were based on classical test theory (CTT) where the test scores are typically obtained as a unit-weighted sum of the item responses. These scores were used in practical applications and in research settings. Increasingly, test publishers utilize item response theory (IRT) in test development and scoring of aptitude and achievement tests. This is mirrored in research applications, where IRT is used to examine these assessment tools. However, for a large number of psychometric instruments in the non-cognitive domain, manifest scores are still used in practical applications – while this has changed considerably in applied research settings. Increasingly, research into constructs such as personality, motivation and work-related attitudes is conducted using structural equation modeling (SEM) where the constructs are specified as latent variables. In contrast to manifest scores, these latent variables are based on weighted items and the unreliability of items is estimated and incorporated in the model. Relationships between constructs can be examined using these latent variables in SEM. Considering these differences between research and practice in generating scale scores of non-cognitive measures, the question arises: To what extend do results obtained based on manifest scale scores differ when compared with results obtained based from latent variables? This question matters as it ascertains whether research results based on latent variable methods can inform practitioners who predominantly assess individuals based on manifest scores. In order to address these concerns, I will first discuss the assumptions of computing manifest versus latent scale scores. Based on simulated data, I will demonstrate the potential differences between both approaches with regard to (a) selection decisions by comparing the rank order of individuals and (b) the impact on the criterion-related validity by comparing the relationships between scale scores and external criteria.

Can we Measure Adolescents' Promotion and Prevention Orientations? Using the Jigsaw Piecewise Technique and Confirmatory Factor Analyses to Answer this Question

Flaviu Hodis[1] and John Hattie[2]

[1]Victoria University of Wellington, Wellington, New Zealand; [2]University of Melbourne, Carlton, Australia

e-mail: flaviu.hodis@vuw.ac.nz

Even when striving to achieve the same goal (e.g., be a good parent), some individuals are driven by ideals and aspirations and have a promotion orientation, while others are motivated to meet their duties and responsibilities and have a prevention focus. As key independent elements of motivation, promotion and prevention have been studied extensively. However, because all extant work has involved college students or older individuals, it is currently unclear whether these constructs can be measured for younger individuals and, if so, how well. This research aims to (a) assess whether younger individuals' promotion and prevention can be measured with a standardized instrument used for more mature respondents; and (b) whether these dimensions are empirically separable. To this end, we employed the General Regulatory Focus Measure (GRFM; Lockwood, Jordan, & Kunda, 2002), an instrument used extensively in motivation research. Our investigation involved two independent samples. In phase 1, 972 high-school students answered the original and unmodified GRFM items. Using Bollen's (2000) jigsaw piecewise technique and confirmatory factor analyses, we identified items that did not perform well and modified them. In phase 2, we administered again the GRFM (including original and modified items) to a second sample of 605 students. With two exceptions, the GRFM items for promotion were appropriate. In contrast, we needed to modify two prevention items and drop four. As the two dimensions had a correlation of 0.99, we also conducted a 1-factor CFA; this unidimensional model fit the data well. High-school students do not perceive promotion and prevention, measured by the GRFM, as distinct constructs. Thus, as developmental stage affects how promotion and prevention can be gauged and the degree to which they are differentiated empirically, an in-depth examination of how other measures function in populations of children and adolescents is necessary.

Setting the Pace: A New Measure of Completion Speed in Item Banked Tests

Tom Hopton

Saville Consulting, Surrey, United Kingdom

e-mail: tom.hopton@savilleconsulting.com

Many measures of how quickly candidates complete tests are inadequate in item banked tests where different candidates are presented with different questions. This paper discusses our newly developed Pace score, which permits a fair comparison of completion rates across candidates who have been presented with different questions. This paper compares Pace with a more conventional Speed score which is based on a count of the number of questions attempted in the time available (e.g. Kurz, 2005). Pace and Speed scores were compared in a live usage sample of N=12,750 individuals who completed the online test Swift Comprehension Aptitude (SCA). Speed was calculated at the overall level, with Pace calculated for the three SCA sub-tests (Verbal, Numerical and Error Checking). There was a substantial ceiling effect in the distribution of the Speed score but not for the Pace scores. Pace differentiates candidates who have attempted all of the questions within the time available, while Speed cannot. When Speed was corrected for this inherent restriction of range, the three Pace scores correlated with Speed in the range of .95 to .97. Pace provides a more appropriate score for use in item banked tests. It is based on a standardised comparison of the time taken by each candidate to complete the specific questions presented to them. Pace is also closely related to Speed, which itself remains useful in

fixed-form tests, particularly where scores are calculated manually. Pace and Speed represent similar constructs, with the advantage of Pace being that it is suitable for use in item banked tests. References Kurz, R. (2005). Convivence of Personality, Motivation, Interest & Ability Theories in Competency. Paper presented at EAWOP Congress, Istanbul, May 2005.

Adding Contextual Information to Interpersonal Empathy in Applied Settings

Miguel Inzunza

Umeå University/Department of Applied Educational Science, Umeå, Sweden

e-mail: miguel.inzunza@edusci.umu.se

Empathy is considered as an important ability in professions dealing with many-sided complex contact situations. With the intention to evaluate differences between groups or individuals, the construct has been measured with different methods and instruments at different occasions. Measures often narrowed to measure empathy at a personal level, interpersonal empathy, where contextual variables have been omitted. These variables might be important when needing to use the interpersonal information in applied settings for example in the police organization. This paper reports the findings of a model of empathy combining interpersonal empathy, as defined in social cognitive neuroscience, with contextual sub-constructs. The contextual sub-constructs incorporate measures of the respondents view on multicultural societies and how often respondents draw inferences based on observed person's appearance. The developed self-report instrument, explores the possibility to incorporate context with an interpersonal measure of empathy (EAI) for future personnel in law enforcement with the objective to develop an instrument more appropriate for applied research in this context. The participants (n=340) are all students in professions dependent on many-sided contact situations with a majority of police recruits (n=176). Other participants are from the economy programme (n=92) and school of education (n=72). A confirmatory factor analysis was used to evaluate the model. Results indicate that there is support for the theorized factor structure. There is also support for convergent validity investigated with a Swedish version of the Interpersonal reactivity index (IRI). These results indicate the value of further development of the instrument. The empathy construct is a complex construct, but to get an applicable model, a measure that incorporates a contextual dimension is required. These first steps suggest more research in the area.

Difference Comparisons of the Primary Grade Students in Computerized Mental Rotation Test

Hi-Lian Jeng and Jhih-Cian Li

Graduate Institute of Digital Learning and Education/National Taiwan University of Science and Technology, Taipei, Taiwan

e-mail: JENGHL@mail.ntust.edu.tw

Gender difference in the performance of mental rotation tasks was explored with manipulations of time constraints and item types. The multiple-choice items in the computerized mental rotation test were developed by arranging systematically the angles to rotate and flip their cubic figures. Primary students of grades 4, 5, and 6 were randomly assigned to receive timed (6 min.) or untimed testing constraint and repeatedly took both types of items (12 items of mirror type and 12 items of different structure type). Results showed that the test can be successfully administered to the young children and find gender differences (in the direction of boy advantage) at grade 4, showing a progress in spatial cognitive development since Linn and Petersen's (1985) meta-analysis and accumulating conforming gender difference to recent research. In the total aggregate as well as in each grade of subjects, no interaction was found but main effects (time constraints, item types, and gender), and there were gender differences in the

scores of different item types by time constraints and grades. The 5th graders outperformed the 6th graders given abundant time to respond (the untimed testing), while the opposite was observed in the timed testing where the 6th graders outperformed the 5th graders. In summary, there existed differences by genders, item types, and grades, and the grade differences were conditioned by time constraint. Multimedia and interactive design should be able to promote future exploration to find if there is even earlier age emerging gender difference and other forms of individual differences in mental rotation tasks.

Psychometric Testing in Forensic Contexts: Developing Standards in the UK

Susan Katherine Jones and Nigel Evans
British Psychological Society, Leicester, United Kingdom
e-mail: susan.jones@psychologyassociates.org.uk

Psychological testing is conducted in a range of forensic practice settings but there were no recognised, formalised set of test user qualifications, unlike Occupational and Educational contexts. Best practice guidelines would suggest that service users be provided with assurance that test users have achieved a suitable level of competence, as important decisions are made based on test performance. The objectives of the study were to develop a set of test user qualifications which lead to registration, that evidence competence in testing with forensic clients and engage psychologists to work towards registration. Design / Method The British Psychological Society (BPS) Committee on Test Standards (CTS) promotes and maintains standards in testing. A CTS working group of eight psychologists from various BPS Divisions and special interest groups was initiated in 2008. Target populations were identified and a model for implementation designed, with ongoing consultations and engagement with a range of stakeholder groups, including companies who sell tests. Results Three certification levels, in line with existing Register of Qualifications of Test Use (RQTU) standards, were developed and approved: Level 1 - Assistant Test User; Level 2 - Test User; and Level 3 - Specialist in Test Use. Modules specifying knowledge and skills for test use in a forensic context have been developed for the first 2 levels. A grand-parenting route for psychologists to the forensic Register has been identified. The development of these standards will assist training course providers in shaping the content of their courses and provide specific competencies on which they can assess their trainees. It will also give psychologists working in forensic contexts an opportunity to develop transferable skills, as well as provide service users with reassurance that test users have achieved a standard of competence.

Increasing Efficacy through Structured Curricula and Immediate Feedback

John De Jong
Pearson / Amsterdam VU University, Velp, Netherlands
e-mail: john.dejong@pearson.com

The success of computer games and the attraction of tablets provide convincing evidence of immediate feedback (IF) and the zone of proximal development (ZPD) as extremely potent means to boost learning. However these principles have yet to be widely and consequently implemented in mainstream educational systems. This is due to a large extent the occurrence of assessment as a post-hoc independent activity and to a lack of understanding how subject matter should be ordered in order to ensure that learners are continuously encountering their next ZPD. To remedy this situation for English language learning a granular scale of increasing functional competence in using English for communication has been defined. Descriptors of language acts have been rated as to their position on this scale independently by a group of about

100 language experts from 11 countries world-wide and by over 300 teachers of English from more than50 countries. A correlation of 0.98 between the two sets of ratings contributes to the validity argument of the ratings. Operationalizing the descriptors an item bank initially containing 2000 items has been developed and field tested with English language learners world-wide. From the items bank tests are constructed that operationalize a two stage assessment with immediate feedback. The first stage is adaptive to quickly find a first preliminary estimate of the area on the scale where the test taker is located. The second part increases the information about the test taker's position reducing the error of measurement and widening the number of aspects of language competence that can be reported. Thus the assessment procedure provides precise and immediate feedback to the learner on how successful their learning is along a continuum of increasing functional communicative ability.

Latent Trait Models for Clinical Skills Performance Examinations: Evaluating Component Skill-Specific Difficulty, Discrimination and Error Variances of Integrated Cases

Nilufer Kahraman[1] and Crystal Brown[2]

[1]Baskent University, Ankara, Turkey; [2]National Board of Medical Examiners, Philadelphia, PA, United States

e-mail: nkahraman6@gmail.com

The United States Medical Licensing Examination Step 2 Clinical Skills examination (USMLE® Step 2 CS) is a standardized patient-based, integrated performance assessment designed to measure the clinical skills deemed necessary for safe and effective supervised practice in the US. The integrated test design of the examination postulates that examinee performance during clinical case encounters is evaluated for four component skills simultaneously: communication, data gathering, spoken English and documentation. A careful study for measurable differences in cases across the content domains of these four components is warranted to highlight situations in which case statistics might be desirable to explain case performance and identify content gaps in CS case pools. This study demonstrates how a latent trait theory- based psychometric approach can be used to collect much needed validity evidence for performance tasks (i.e., CS cases) that can help evaluate clinical content domain ties of individual CS cases. The following section discusses current methods for evaluating the performance of integrated cases (the SP-focused approach) followed by a practical argument for a different (task-focused) approach. A set of linear latent trait models were fit to a sample of 2,284 examinees who took the CS examination in 2010 to obtain component skill-specific case difficulty, discrimination and error variances. Linear latent trait modeling appears to be a promising tool for obtaining component-specific case statistics. The results suggest that difficulty and discrimination parameter estimates may help quantify content-specific properties of CS cases and identify weaknesses and strengths of the overall CS case pool. Availability of component-specific case statistics would help monitor performance of cases across clinical content domains and spot content gaps in the CS case pool to tailor future case development efforts and ensure comparability of exam forms. They may also uncover psychometric properties of varying assessment formats.

The Classroom Assessment Standards: Guidelines for Teacher Practice

Don Klinger

Queen's University, Kingston, Ontario, Canada

e-mail: don.klinger@queensu.ca

While research has largely focused on large-scale testing in k-12 education, classroom teachers conduct the vast majority of testing (assessment) activities. Given the amount of testing teachers conduct, the Joint Committee on Standards for Educational Evaluation sponsored the

development of the Classroom Assessment Standards (CAS). The CAS contain a set of ANSI approved Standards and related guidelines approved by the 17 professional organizations (e.g., AEA, AERA, APA, CES, CSSE, NCME) within the JCSEE. The standards and guidelines identify the foundations for sound classroom assessment practices, and the issues to be considered when conducting formative, diagnostic, or summative assessment, testing, and decision making for all students. Classroom assessment practices that adhere to these standards can be used with confidence by teachers and where appropriate, students, to better foster student learning. The intentions of the CAS are to: 1) provide standards for practice to guide classroom teachers' assessment practices; 2) develop teachers' confidence and skills related to research supported assessment practices; 3) identify the roles, purposes and audiences for classroom assessment; and 4) describe assessment practices that fairly support the learning of all students. The CAS were developed by a task force selected by the JCSEE (www.jcsee.org). For the past four years, the task force has used the research literature and other relevant standards to develop the CAS. The development process has included wide dissemination of the CAS drafts and consultation with members from each of the JCSEE member organisations, field trials, national and international hearings, and finally a expert validation Pannel. This paper session will present the recently approved Classroom Assessment Standards, illustrating how the CAS support sound testing and assessment practices in today's K-12 classrooms. Session participants will also be invited to provide further input into the continual revision process for the standards.

Understanding Preservice and Inservice Teachers' Assessment Literacy and Conceptions of Assessment

Kim Koh and Dave Scott

University of Calgary, Calgary, Alberta, Canada;

e-mail: khkoh@ucalgary.ca

The implementation of assessment for learning practices in schools cannot be realized if teachers do not perceive that improved teaching and student learning are the true functions of assessment. Teachers' conceptions of assessment tend to be ecologically rational; that is, aligned with the social, cultural, and policy priorities of their school context. Therefore, when performativity on high-stakes, standardized testing is a priority of both school and system, assessment is likely to be perceived as mainly serving accountability functions. Assessment for learning practices can also be impeded if teachers have a low level of assessment literacy. The purpose of this study is to examine preservice and inservice teachers' assessment literacy and conceptions of assessment in a high performative culture. Approximately 300 Canadian teacher candidates and 250 teachers were administered the Assessment Literacy Inventory (ALI) and the Teachers' Conceptions of Assessment abridged inventory (TCoA-IIIA). The ALI is adapted by the authors for use in the Canadian context. It consists of 42 multi-choice items to measure preservice and inservice teachers' levels of assessment literacy. The TCoA-IIIA consists of 27 6-point Likert items, which measure four major conceptions of assessment: assessment improves teaching and learning, assessment makes students accountable, assessment makes teachers and schools accountable, and assessment is irrelevant. Confirmatory factor analysis will be used to validate the factor structures in the data. The relationship between teachers' assessment literacy and conceptions of assessment will be examined using structural equation modeling. The data analysis will be completed for our presentation at the ITC conference in July 2014. We expect that the results will inform the planning and review of teacher education and professional development programs in Canada and countries that share similar reform visions in assessment.

Additionally, the psychometric properties of the adapted version of the ALI will be discussed at the conference.

The Assessment of Collaborative Problem Solving: The Approach of The Experiment-Based Assessment of Behavior

Katarina Krkovic and Samuel Greiff
University of Luxembourg, Luxembourg
e-mail: katarina.krkovic@uni.lu

New trends in the 21st century labour market foresee people to be able to work together effectively. Not only are people expected to work in teams, but also to be able to solve complex, new problem situations in the course of such collaboration. Due to the fact that collaborative problem solving skills are becoming a common requirement at the workplace and in everyday life, psychologists and educationalists alike have started emphasizing the importance of measuring and fostering these skills throughout life. We developed an innovative assessment instrument measuring collaborative problem solving. This instrument can be categorized as an experiment-based assessment of behavior sensu Cattell. It captures collaborative behavior of the test-taker in multiple complex problem solving tasks. Complex problem solving tasks within our instrument are based upon the MicroDYN approach, a typical computer-based micro-world approach that has been extensively validated for individual problem solving. The instrument uses computer-simulated agents as collaborators, offering standardized collaboration stimuli to each test-taker. In the first pilot study, we administered the instrument to 92 12-year-old students in one school in Germany. First results suggest that the instrument functioned as anticipated, and indicate good construct validity. Moreover, the results show that students manifested stable collaborative problem solving behavior across diverse problem situations, whereby students actively interacted with the computer-agent and worked on solving the problem. Additionally, students differed in using specific patterns of collaborative behavior as expected. In the discussion part we connect the theoretical foundations of collaborative problem solving with behavioral patterns found in our study. Moreover, we discuss general psychometrical constraints of assessing collaborative problem solving in a standardized manner. Furthermore, we elaborate on using computer agents in assessments, and upcoming developments in our experiment-based behavior assessment of collaborative problem solving.

Cognitive Diagnostic Models for the Computerized Test with Multiple Choice and Constructed Response Items

Bor-Chen Kuo[1], Huey-Min Wu[2], Shu-Chuan Shih[3], Chun-Hua Chen[3] and Hungsheng Lin[3]
[1]National Taichung University of Education, Taichung City, Taiwan; [2]Research Center for Testing and Assessment, National Academy for Educational Research, New Taipei City, Taiwan; [3]Graduate Institute of Educational Measurement and Statistics, National Taichung University of Education, Taichung, Taiwan;
e-mail: kbc@mail.ntcu.edu.tw

Traditionally, multiple choice items are widely used in computerized tests. But constructed response items that elicit students' higher-level constructs are beneficial to evaluate complex concepts or skills such as procedure knowledge or problem solving. For reliability and validity consideration, the scoring process of constructed response items is time-consuming, and with expensive human-resource. In this study, an automated scoring mechanism is introduced for constructed response items to diagnose the status of concept and misconception during problem solving process. Most of cognitive diagnostic models are designed for tests with multiple choice items to reduce the uncertainty of attribute estimates. Traditional cognitive diagnostic models

are not enough to handle the richness of information provided by constructed response items. For cooperating multiple choice and constructed response items simultaneously, some novel cognitive diagnostic models are proposed in this study. A test of "multiplication and division of fractions" based on the 5th grade mathematics curriculum in Taiwan was developed to evaluate the proposed cognitive diagnostic models. The experimental results show that :1)The estimated student skills and bugs from the proposed automated scoring mechanism has a consistency rate (above 98%) with the judgment of domain experts;2)The proposed cognitive diagnostic models can deal with multiple choice and constructed response items simultaneously and outperform traditional cognitive diagnostic models.

The Bilingual Assessment of Cognitive Abilities

Serge Lacroix
University of British Columbia, Vancouver, Canada
e-mail: auguston@shaw.ca

The number of bilingual individuals encountered in the school systems is constantly increasing to the point where the majority of the student population in all important urban centres in North America is now comprised of more than 50% of people speaking a language other than English at home. Language minority students tend to be over-represented in special education categories and placement. In this study the role that language plays in the expression of intelligence, bilingualism, and the process of assessing selected cognitive abilities was explored. The primary purpose of the study was to determine if individuals who are allowed to move from one language to another when they provide responses to test items produce results that are different than those obtained by bilingual examinees assessed in one language only. Bilingual students were tested in both English and French on various measures of cognitive abilities. The results indicate that the Experimental Group obtained significantly higher results than the Control Group on all the tests and subtests used. The Experimental Group code-switched (moved from one language to the other) more frequently and the examiners only code-switched with that group. The frequency of the code-switching behaviours explains, in great part, all the differences noted in the results as very few other sources of differences were identified, even when groups were compared on sex, first language and relative proficiency in French and in English. Given these results, a new testing procedure is suggested to better address the needs of the multilingual population and account for the use of their languages during testing.

Investigating SES-Related DIF (Differential Item Functioning) across Countries for PISA Mathematics Items*

Luc Le
Australian Council for Educational Research, Camberwell, VIC, Australia
e-mail: luc.le@acer.edu.au

Socio-economic Status (SES) is an important explanatory factor in many studies in health, psychology, child development and education. As such, it is often a strong predictor of academic achievement. This study uses data from PISA 2003 (for which Mathematical Literacy was the major focus) to explore the effect student SES had on performance on the mathematics items, which focused on real-world topics. An IRT method was implemented to detect uniform Differential Item Functioning (DIF) between students with lower- and higher-SES backgrounds in each of 41 PISA countries. Relative weaknesses and strengths of these student groups were identified with respect to the classification of the items in the PISA framework by content, context, competency and format. Item format was consistently detected as the most significant

factor for the SES-related DIF in the PISA countries. In most countries, lower-SES students tended to be most disadvantaged on open constructed-response items and advantaged on multiple-choice, complex multiple-choice and closed-constructed response items. There was no significant effect of item content, context and competency on SES-related DIF in most countries. However, it is likely that lower-SES students are most disadvantaged on Change/Relationships (content), Public (context), and Reflection (competency); and most on Quantity (content), Personal and Educational/Occupational (context), and Connections and Reproduction (competency). Findings from this study provide a potentially valuable contribution to the development of national and international large-scale testings or assessments.

Priming Effects of Information Sources on Escape Judgment

Hong Li
Tsinghua University, Beijing, China
e-mail: mumumaohuang@gmail.com

This article aims to explore the priming effect of information sources on escape judgment. Based on this purpose, we conduct 3 different experiments that describe the effects of information sources on judgment based on different escape choices. Experiment 1 and 2 adopt a 3 (information sources: expert, acquaintance, stranger) X 2 (escape choice: following crowds, individual action) within-subjects design, and the dependent variable is escape judgment. Experiment 3 uses a 2 (situation: fire, common) X 2 (escape choice: following crowds, individual action) within-subjects design, and the dependent variable is preference for escape choice. The main findings are: (1) Under unconscious priming condition, the main effect of information sources on escape judgment is very significant, and the main effect of escape choice on judgment is also very significant. (2) Under conscious priming condition, the main effect of information source on escape judgment is very significant, and the main effect of escape choice on judgment is very significant, there is a significant interactive effect between information source and escape choice on escape judgment. Specifically, when the priming word is expert, the difference between following crowds and individual action is not significant. However, when the priming words are acquaintance and stranger, the difference between the two escape choices is very significant. (3) There is neither main effect of situation nor escape choice on judgment reaction time. There is also no interactive effect between situation and escape choice on judgment reaction time. In conclusion, judgment on escape choice is significantly affected by information sources.

The Interaction between Rater's Regulatory Focus and Applicant's Impression Management Strategy in Selection Interview: From Regulatory Fit Perspective*

Zheting Lin, Ran Bian, Hongsheng Che and Qin Gao
Beijing Normal University, Beijing, China
e-mail: linzheting@126.com

Researchers have showed that individual's strategy of goal pursuit can be influenced by others whose strategy is perceived to fit(vs. not fit) his or her orientation. Specifically, in social process, people with promotion focus is easier to accept eagerness strategy, and prevention individual is more likely to accept vigilance strategy. In fact, the interaction between rater and applicant in selection interview is typically an interpersonal exchange process. So this study made the first attempt at investigating whether such interpersonal regulatory fit exits between the evaluation goal of rater and impression management strategy of applicant. Two experiments were conducted to exam the regulatory fit effect between rater's regulatory focus and applicant's

strategy. Moreover, the mechanism of the interaction and prospects for follow-up study were discussed. Both experiments adopted 2 (regulatory focus of rater: promotion vs. prevention) ×2 (impression management strategy of applicant: assertive vs. defensive) between-subject design. Experiment 1 used verbal video-based interview materials and experiment 2 used nonverbal video-based materials. Subjects were first trained to be an interviewer, then they were asked to evaluate the applicant's performance and likability. Data were collected from 238 college students (118 in Experiment 1and 120 in Experiment 2). Taken together, Experiment 1 revealed significant interaction between rater's focus and applicant's verbal impression management strategy. Experiment 2 showed the regulatory fit also exited between rater's regulatory focus and applicant's nonverbal IM strategy. Specifically, rater with promotion focus evaluated higher on applicant with assertive IM strategy, and liked them more. In contract, rater with prevention focus liked applicant with defensive IM strategy more and gave them higher score.

The Context Effects of Preferences for Intuitive and Analytic Decision Strategies

Alice Huang Linyan
Tsinghua University, Weiqing Building, Beijing, China
e-mail: alice.huang@163.com

When people confront with different alternatives, one thing they can do is rely on their intuition, that is, base their decision on the first affective reactions that come to mind. They could also decide deliberately, analyzing the pros and cons of different options before making a decision. Although historically, philosophers and other scholars have stressed the importance of (logical) reasoning to achieve rationality, and intuition has been blamed for a range of cognitive biases in the psychological literatures on reasoning and decision making for a long period. Fortunately, due to Simon's seminal perspective of "Bounded Rationality" and "Emotion Revolution" in cognitive psychology which was raised at the end of last century, researchers in the domain of decision making have recognized that, intuitive and analytic decision making both occupy important and irreplaceable positions. The purpose of the present study was to examine how the context influences individual strategy preferences. We conducted two experiments on the explicit and implicit levels respectively, investigating to what extent people tend to adopt intuitive or deliberative decision strategies under the situations of emergency and consumption. Main findings are listed as below. (1) On both of the explicit and implicit levels, people preferred to an intuitive strategy under emergence situation. (2) On both of the explicit and implicit levels, people preferred to deliberative strategy under consumption situation. (3) People more tended to use deliberative strategy than intuitive strategy no matter under either of the situations. The conclusion is that the context of emergency is more likely to prime intuitive strategy, while the context of consumption tends to prime deliberative strategy.

The Development of Forced-Choice Personality Scale for Neuroticism

Hongyun Liu, Hui Li, Dong Zhang and Fang Luo
School of Psychology, Beijing Normal University, Beijing, China
e-mail: hyliu@bnu.edu.cn

Using personality tests in personnel selection becomes increasingly popular in industries in China. The validity of personality tests with Likert scale is challenged due to the possibility of getting faking answers. Forced-choice scales are often used to deal with faking. The scores of forced-choice scale are usually ipsative data, while MDPP (multidimensional pairwise preference) model focuses on the mathematic model of subjects' response to Forced-choice scale, which can

more accurately estimate the traits and provided new insights into the problem of ipsative data in last several years. This research developed the Neuroticism Forced-choice Scale and Likert scale, which were administrated simultaneously in three conditions (stress-free condition, real employment and simulated employment conditions). The scales included five dimensions, which are Anxiety, Anger/Hostility, Depression, Self-awareness, Impulsivity, and Vulnerability. Furthermore, MDPP model was expanded, and its accuracy was testing in this study. The results demonstrated that the data from administrating the Neuroticism Forced-choice Scale well fitted the expanded-MDPP model than MDPP model. The traits' scores estimated by expanded-MDPP model were normative data, and better resisted faking than the scores gotten in the traditional way. Furthermore, forced-choice Scale better resisted faking than the Likert scale.

Differential Item Functioning with Respect to Test Takers' Culture Background

Jinghua Liu[1] and Tim Moses[2]

[1]Secondary School Admission Test Board, Skillman, United States; [2]Educational Testing Service, Kent, WA, United States

e-mail: jliu@ssat.org

Differential item functioning (DIF) examines whether an item performs invariantly across different subgroups given the same abilities. It has been implemented as a fairness procedure for more than two decades. The decision regarding which subgroups need to be examined for DIF depends on the testing population. When a testing population is homogeneous, gender DIF may be sufficient. When a testing program evolves and the population becomes more heterogeneous, DIF practice needs to evolve as well in a way that warrants DIF on different subgroups in the current population. Test takers' culture background may have non-trivial influence on item performance. The purpose of this study is to examine DIF from the perspective of test takers' culture background: U.S culture vs. Non-U.S culture. Data Data are collected from the Secondary School Admission Test (SSAT). The SSAT measures verbal, quantitative and reading skills students develop over time. It is designed for students who apply to private schools in the US and Canada. When the SSAT was launched in the 1950s, the test taking population was US domestic solely. Over the years, more and more international students take the SSAT. In 2012-13, more than 20% of the test takers were international students. There might be potential DIF problems for international students. For example, an item regarding feet and inches might be differentially harder for Asian and Canadian students who use meters and centimeters in their system of measurement. Analysis Plan The reference group includes U.S domestic students, whereas the focal group includes international students. The Mantel-Haenszel (MH D-DIF) statistic and standardized P-differences (STD P-DIF) statistic will be used as DIF statistics. We hope that this study sheds lights on item writing and test development (e.g., avoids the use of culture-specific items) for any testing programs that are oriented globalization.

Online Assessment Meets Serious Gaming

Annette Maij-De Meij and Lolle Schakel

PI Company, Utrecht, Netherlands

e-mail: a.de.meij@picompany.nl

Recruitment strategies have fundamentally changed over the past decade. The arrival of the web has had a substantial impact on organizations' recruitment and selection methods. Implementing online selection procedures which help increase the number of applications which match perfectly with the advertised vacancies and reduce the number of unsuitable incoming applications is a major challenge. An important reason for the large number of unsuitable

applications is that many candidates have unrealistic ideas as to what a particular job actually is or requires. Internet based game-like applications can help meet this challenge. A game that consists of job relevant challenges can perform two important functions: it can give both a realistic impression of a typical work day for a candidate and function as a pre-selection device for the organization. The study describes the development of a game-like online environment for the selection of call center employees. The online environment is based on standard assessment tools like personality questionnaires, cognitive capabilities tests, SJTs, language skills, computer skills etc. In the online environment the candidate is taken through different phases. The candidate is provided with information about the organization and is confronted with customers calling for which a number of tasks need to be performed that reflect the job requirements. Furthermore, by an online dashboard the recruiter is supplied with an automatic ranking of candidates that meet the requirements, while the candidates can compare their results with the performance of successful call center employees. The construction of the different parts of the online assessment will be discussed and examples will be shown. A remarkable cost-reduction in the recruitment of candidates and decrease of turnover with 20% have been accomplished. An evaluation of applicant reactions and results of predictive validity studies will be presented as well.

Easily too Difficult: Estimating Item Difficulty of Microworlds

Stadler Matthias and Greiff Samuel
University of Luxembourg, Luxembourg
e-mail: matthias.stadler@uni.lu

Dealing with complexity and dynamics is more and more becoming part of people's everyday lives. This increased necessity to deal with complex systems has instigated different fields of research utilizing computer simulations, so-called Microworlds (MWs), to observe and study human behavior in complex situations. While these MWs enjoy great popularity with both practitioners and researchers, their psychometric qualities have been questioned and studies investigating those qualities are sparse. The aim of this study is to explore the factors contributing to MWs' item difficulty using multiple small independent MWs. We analyzed data of 2965 Finnish school children, using a linear logistic test model. In this model, item difficulty is predicted by item characteristics rather than estimated from the data. Based on previous research, we chose six relevant item characteristics, which were varied over nine independent MWs. Our results show that MWs' item difficulty is almost perfectly predictable by six basic characteristics. These results are consistent over two independent sub samples. We conclude that even small MWs can be very difficult depending on the presence and amount of those characteristics. This raises the question whether large MWs, which are frequently used both in practice and research, are too difficult for most participants and thus do not provide reliable and valid estimates of behavior or skills. In addition we provide evidence for the utility of differentiating between difficulty of controlling a MW and understanding it's underlying system. Finally, we discuss limitations of our research as well as further theoretical and practical implications.

Models of Equity: An International Comparison of The Relationship between Schools' Average Economic-Socio-Culture Status and PISA Mathematics Achievement

Carina McCormick[1], Leslie R. Hawley[2] and Kurt F. Geisinger[3]

[1]University of Nebraska-Lincoln, Lincoln, NE, United States; [2]The Nebraska Center for Research on Children, Youth, Families & Schools (CYFS), Lincoln, NE, United States; [3]Buros Center for Testing, Lincoln, NE, United States

e-mail: mccormick@huskers.unl.edu

As countries seek to ensure high-quality educational opportunities for all students, the connection between students' economic-socio-cultural status (ESCS) and their achievement is a primary concern. In many high-performing countries on the Programme for International Student Assessment (PISA) in 2012, this relationship was smaller than the OECD average, suggesting greater equity is associated with higher overall achievement (OECD, 2013). The current study considers the unique relationship between a school's average ESCS and achievement, after accounting for other characteristics of the school (such as classroom culture, discipline, and motivation), and determines how the strength of this relationship differs between countries. The researchers will estimate a three-level model (i.e. students, school, and countries) using the HPMIXED procedure in SAS 9.3, incorporating several non-economic features of school quality as well as students' ESCS, with mathematics achievement (plausible values) as the outcome. The focus of results will be on inclusion of random effects by country for school mean ESCS, which produces a unique estimate of the relationship between it and student achievement for each country. Such values are not typically interpreted but are used similarly in procedures such as value-added modeling. Preliminary results showed that although there was not a strong fixed effect of a country's mean ESCS level on the country's achievement, within the United States there was a strong effect of a school's mean ESCS on the school's achievement, supporting the need to statistically examine country differences in this relationship in new ways. Because results will isolate the effect of a school's socio-economic composition in a larger model, results will more clearly support direct interpretation regarding between-country differences in its relationship to student achievement. This relatively novel application of multilevel modeling features can also be used with other variables to examine how effects vary across countries.

The Assessment of Fatigue: A Comparison between Three Instruments

Enrique Merino-Tejedor

University Of Valladolid, Segovia, Spain

e-mail: enmerino@psi.uva.es

The assessment of fatigue is an important issue since it has important outcomes on well-being and performance, so it is important to develop instruments in order to get an accurate assessment of this construct. The main objective of this study was to compare the assessment of fatigue through three instruments to deepen on the nature and the dimensions of this concept. A multi-occupational Spanish sample of 364 subjects participated in this investigation. Spanish versions of the following three instruments were used: The Swedish Occupational Fatigue Inventory (SOFI) (Ahsberg, 1998), the Work Effort Scale (WESC) (De Cooman, De Gieter, Pepermans, Jegers, and van Acker, 2009), and the Fatigue Assessment Scale (FAS) (Michielsen, De Vries, van Heck, van de Vijver, and Sijtsma, 2004). The SOFI consists of five factors: lack of energy, physical exertion, physical discomfort, lack of motivation, and sleepiness; the WESC consists of three factors: persistence, direction, and intensity; finally, the instrument FAS offers a global score of fatigue. We obtained reliability analysis and correlation analysis. Among the results obtained we found sound reliability values for the three instruments, alpha coefficient of

.933 for the SOFI, .751 for the FAS, and .924 for the WESC. It must be taken into account that in the WESC a high score indicates a lower fatigue, so it was expected a negative correlation among the WEC and the other two instruments. Significant correlations were found between global scores of the three instruments. However, mixed results were found among the dimensions of the instruments, for example, no significant correlations were found between the component intensity of the WESC and most of the SOFI components. Among the of this study, it must be considered that fatigue is a complex concept and some of the dimensions assessed by current instruments offer different results to be considered.

Accounting for Wording Effects in the Adaptation of a Scale for Basic Psychological Needs

Michalis Michaelides[1], Kyriaki Fousiani[2] and Panayiota Dimitropoulou[1]

[1]University of Cyprus, Nicosia, Cyprus; [2]Neapolis University Pafos, Pafos, Cyprus

e-mail: michalim@ucy.ac.cy

Basic Psychological Needs Theory, a part of Self-Determination Theory (Deci & Ryan, 2000) proposes that autonomy, relatedness and competence are universal needs essential to psychological health and well-being. Satisfaction of the three needs is linked to wellness, while frustration with each relates to ill-functioning and pathology. A number of scales have been developed to measure the satisfaction and frustration aspects of the three basic needs. The study examined the factor structure of the adaptation of van Petegem et al.'s scale in Greek after considering the presence of method effects due to wording. Data were collected from 545 participants in Cyprus on 24 items on autonomy, relatedness and competence needs. Half of the items on each subscale were positively worded and measured need satisfaction; the remaining items were negatively worded and measured need frustration. Eleven confirmatory factor analysis models were fitted using AMOS 20.0 to evaluate the best fit. Two categories of models were employed to account for both substantive and wording effects: Correlated Trait Correlated Uniqueness (CTCU) and Correlated Trait Correlated Method (CTCM). A model where only three substantive needs factors were specified was not acceptable. Among the CTCU models, the best fit was achieved when all negatively-worded items had correlated error variances. CTCM models were more parsimonious according to the Corrected Akaike Information Criterion; a model with three substantive, three positive and three negative factors had the best fit, followed by a model with three substantive, one positive and one negative factor. After accounting for wording effects, the scale demonstrated three clear substantive dimensions as hypothesized by the Basic Psychological Needs Theory. Including method factors allows for further examination of the nature of "positive" and "negative" latent variables. Moreover, it appears that the wording effects may be slightly different for each substantive factor.

The Questionnaire Competitiveness in Management (Approbation and Developing)*

Olga Mitina[1], Nina Nizovskikh[2] and Marina Sharafutdinova[2]

[1]Lomonosov MSU, Moscow State University of Psychology and Education, Moscow, Russia; [2]VSHU, Kirov, Russia

e-mail: omitina@inbox.ru

The method which allows to measure the severity of the basic personality traits that determine competitiveness in the field of management (the questionnaire psychological readiness toward managing was recently developed (Tiulkin, Nizovskikh, Mitina 2012). The aim is to identify common predisposition to management and the dominant attitudes in this area. Theoretical basis of the test is the concept of human life principles (Nizovskikh) and manager competency model L. and S. Spencers (2005).The names and the sequence of 15 scales of the test match

competencies manager identified on the basis of empirical studies involved big samples: 1.Ability to impact and influence; 2.Focus on achieving; 3.Ability for teamwork and cooperation; 4.Strong analytical thinking; 5.Initiative; 6.The ability for developing other; 7.Self-confidence; 8.Interpersonal understanding; 9.Prescriptive / Perseverance; 10.Desire to seek new information; 11.Intensity of team leadership; 12.High level of conceptual thinking; 13.Understanding of the company and building relationships; 14.Caring for the order; 15.Technical awareness. Each item of the questionnaire is a principles of the life (totally 150 items). Respondents were University students different genders both gender living in Moscow and regional center Vyatka with different levels of experience in managing groups of people. Empirical data confirmed high level of reliability and correlation validity. These parameters with a predominance of certain styles and features of self-management and self-regulation (J.Kuhl), time perspective structure (Ph.Zimbardo), basic values (S.Schwartz) were used for correlation analysis. Aslo we differentiated the acceptance this or that principle (as declaration) and real following to it. So each subject should to answer each item twice: • in what degree "I want to do it" and • in what degree "I manage to do it". Two different sets of personal scales each of which corresponds to own modality were got. The analysis of differences between these two modalities also is useful.

When Anchor Items are Administered under Different Conditions: Modeling Differences In Motivation

Marie-Anne Mittelhaeuser[1], Klaas Sijtsma[2] and Anton Béguin[3]

[1]Tilburg University/ Cito, Elst, Netherlands; [2]Tilburg University, Tilburg, Netherlands; [3]Cito, Arnhem, Netherlands

e-mail: M.Mittelhaeuser@uvt.nl

Linking procedures are used to disentangle differences in test form difficulty and differences in the proficiency of examinees so that scores on different test forms can be used interchangeably. An effect that needs to be taken into account when choosing a data collection design for linking purposes is 'differential motivation', which refers to a situation where an examinee might work harder on an item in a high-stakes administration condition compared to a low-stakes administration condition. We will discuss different data collection designs for linking purposes and their robustness against the effect of differential motivation. Furthermore, theoretical considerations for choosing one type of data collection design over the other with respect to differential motivation, will be supported with results from (1) a simulation study and (2) an empirical example. The effect of differential motivation and the robustness of linking procedures will be discussed for the non-equivalent groups design, the common-item non-equivalent groups design, the pre-test design and the post-test design. In the simulation study performance of IRT linking based on the Rasch model and IRT linking based on the mixture Rasch model is compared for the common-item non-equivalent groups design and the pre-test design. The pre-test design is the most vulnerable to the effect of differential motivation, and will result in an overestimation of the differences in proficiency between the two populations administered the high-stakes test forms to be linked. The effect differential motivation might have on the linking result should be taken into account when choosing a data collection design for linking purposes. Furthermore, finding methods to gain more insight in the effect differential motivation has on linking high-stakes tests would be valuable for measurement practice and measurement research.

A Cross-Cultural Investigation of The Structure of Holland's Personality Type Model in South Africa

Brandon Morgan and Gideon De Bruin
University of Johannesburg, Johannesburg, South Africa
e-mail: bm.morgan@yahoo.com

Holland's circular/circumplex theory of vocational personality types (Realistic, Investigative, Artistic, Social, Enterprising, Conventional) is recognized as one of the most influential vocational counselling theories. Despite its widespread use, there remain questions around its cross-cultural transportability, particularly to countries outside of the United States of America. Research conducted in South Africa has found poor model fit across different cultural groups. This is problematic because career counselling based on his theory is moot if the structure is not supported. Accordingly, the purpose of this paper is to investigate Holland's structural model across different cultural groups in South Africa. The objectives of this study were 1. to investigate the fit of Holland's circular/circumplex model in different cultural groups; and 2. to compare the fit of the models across different cultural groups. A total of 985 participants from different cultural contexts completed the South African Career Interest Inventory. The data was analysed using unconstrained and constrained multidimensional scaling (MDS) and the randomisation test of hypothesised order relations. Unconstrained MDS demonstrated the correct ordering of the six types across the different cultural groups while constrained MDS found that the data could account for much of the variance in the model. The RTHOR found that the model fit was good and that there were no differences in the fit when competing models were compared with each other. The results suggest that Holland's circular/circumplex model does fit across different cultural groups and that there are no significant differences in the fit when compared to each other. Therefore career counselling based on Holland's theory may proceed in the South African context.

Different Performance at Test End by Age and Gender in an Aptitude Reasoning Test with Multiple–Choice Items

Van Nguyen
Australian Council for Educational Research, Camberwell VIC, Australia
e-mail: van.nguyen@acer.edu.au

The Special Tertiary Admissions Test (STAT) was developed by the Australian Council for Educational Research (ACER) at the request of the Australasian Conference of Tertiary Admission Centres (ACTAC). The test has been used since 1992, as an alternate method of gaining entry into Australian university courses, for people who do not hold a recent Year 12 certificate in Australia. The test compromises 64 scored multiple-choice items, half of which are Verbal Reasoning questions and the other half, Quantitative Reasoning questions. This study uses 2012 STAT data, comprising about 12,000 candidates, 55% of which were female and 45% of which were male. This study was conducted to explore speed effect on different gender and age groups (20 or less, 21-24, 25-30, and more than 30). In each domain, we grouped sequence items into four clusters (8 items each) and investigated the success of performance groups in each cluster, together with a rate of missing responses. Rasch item bundle fit statistics were also employed using ConQuest software to examine the fit of the clusters to the IRT Rasch model. Main results show that in both domains, the last cluster had the lowest fit to the Rasch model. Moreover, males tended to make slightly more effort and perform more successfully than females towards to the test end. Young candidates tried to answer more questions in the last clusters and they

performed slightly better than older candidates at the end of the test, especially within the Quantitative Reasoning domain. Results from this study would be useful for future test development or such studies related to test or item difference performance, by gender or age.

Construct Validity of the Construction Task

Becker Nicolas[1], Florian Schmitz[2], Anke Falk[1], Daniel Recktenwald[1], Jasmin Feldbrügge[3], Franzis Preckel[3], Oliver Wilhelm[2] and Frank M. Spinath[1]

[1]Universität des Saarlandes, Saarbrücken, Germany; [2]Universität Ulm, Ulm, Germany; [3]Universität Trier, Trier, Germany

e-mail: nibe1@gmx.de

Figural matrices tasks are classical markers for abstract reasoning ability. The Construction Task (CT) is a novel item format for the presentation of figural matrices. Test takers have to generate the correct answer in a computerized testing environment instead of choosing it out of a given set of response options. This approach avoids "distractor weakness", i.e. the problem, that some test takers pursue a response elimination strategy. This strategy involves identifying as many rules as possible, eliminating response options regarded as wrong, and guessing among the remaining ones. Reliability as well as factorial and criterion validity of the CT have been demonstrated in previous studies. The present study deals with the construct validity of the CT in relation to Working Memory Capacity (WMC) which is closely related to abstract reasoning. Since the CT prevents differential solution strategies, WMC might be more relevant than in the usual item format. To put this to the test, we compared the strength of associations between a working memory test battery and two matrices tests: one presented in the conventional format (i.e., in using distractors) and one presented as a CT. A total of 101 participants took a WMC test battery. Group 1 consisted of 54 participants who also completed 38 items presented as CTs. Group 2 included 47 participants who completed the same 38 items in the conventional format. We found a substantially higher correlation between WMC and the items presented as CTs (r = .52) than between WMC and the items presented in the usual format (r = .27). The findings confirm our hypothesis that WMC is more relevant in the CT than in the usual format. Implications for the construct validity of the CT will be discussed.

The Validity of Forced-Choice Personality Measures in Operational Testing Environments

Christopher Nye[1], Fritz Drasgow[2], Leonard White[3], U. Christean Kubisiak[4], Oleksandr Chernyshenko[5] and Stephen Stark[6]

[1]Michigan State University, East Lansing, MI, Michigan, United States; [2]University of Illinois, Champaign, Illinois, United States; [3]U.S. Army Research Institute, Fort Belvoir, Virginia, United States; [4]PDRI, a CEB Company, Tampa, Florida, United States; [5]Nanyang Technological University, Singapore, Singapore; [6]University of South Florida, Tampa, Florida, United States

e-mail: nyechris@msu.edu

Despite positive findings demonstrating the validity of personality assessments for employee selection, faking continues to be an important concern for operational employment tests. As a result, recent efforts have attempted to address this issue using forced-choice personality measures. The goal of the present study was to examine the predictive validity of a forced-choice personality scale in a high stakes setting. The Tailored Adaptive Personality Assessment System (TAPAS) uses a multidimensional pairwise preference (MDPP) format which was designed to be resistant to faking. In the past, forced-choice response formats produced only ipsative scores (i.e., scores that sum to the same constant for each respondent), which are largely unsuitable for personnel selection. However, the MDPP scoring system used for the TAPAS has overcome this

major limitation and is capable of recovering normative scores regardless of how many dimensions are assessed. The data for this research included TAPAS and criterion data collected by the U.S. Army from a total of 151,625 enlisted Soldiers who were administered the TAPAS as part of the application process. Those that were selected into the Army then entered training and criterion data was collected at the end of the training program. Results showed that the TAPAS scales were significant predictors of performance outcomes and attrition (multiple R's ranged from .14 to .33 across all MOS). Importantly, individual facets were not as predictive as composites of the TAPAS scales. In addition, preliminary results also indicated that the pattern of relationships among the TAPAS scales and criterion composites differed across occupations, suggesting that person-environment fit may be particularly important for establishing the validity of personality assessments. These results provide some initial evidence that forced-choice personality measures may be useful for employee selection, even in high-stakes operational settings.

Effective Test Score Reporting: Improving Test Validity through Establishing Evidence Based Design Principles for Test Score Reports

Tim O'leary[1] and John Hattie[2]

[1]University of Melbourne, Melbourne, Australia; [2]University of Melbourne, Carlton, Australia

e-mail: olearyt@students.unimelb.edu.au

Whilst there are many elements that make-up validity this paper aims to explore the concept of one critical aspect – the appropriateness of interpretations that various stake holders make based on test score reports. The purpose of this paper is to investigate the design features associated with the construction of test score reports with a view to identifying those features that lead to both valid and invalid inferences being made. The aim is to develop a set of over-arching design methodologies and principles for the construction of test score reports that lead to the appropriateness of interpretations and inferences. This paper presents a preliminary set of design principles based upon a contemporary literature search and an initial study, using "think-aloud" protocols of various educational stake-holders in order to determine exactly what it is about test score reports that lead to valid inferences being made.

Pioneering the Use of Natural Language Processing Tools to Enhance the Analysis of Sources of Differential Item Functioning in Assessments Administered Internationally*

Maria Elena Oliveri, Frederic Robin, Rene Lawless and John Young

Educational Testing Service, Princeton, NJ, United States

e-mail: moliveri@ets.org

Central to developing fair cross-cultural assessments is ensuring that scores yield the same meaning across populations. Differential item functioning (DIF) analyses are typically conducted for this purpose. DIF occurs when examinees of similar ability levels from different populations perform differently on a test item, suggesting differential measurement of the construct, threatening the score consistency and test validity. In DIF analyses, the probability of correctly responding to an item is calculated for a focal group (the minority group) relative to a reference group (for whom the item is advantaged). After DIF detection, expert reviews are conducted to examine potential item bias to inform decisions about whether to delete or revise an item to minimize biases in the item pool. Items with large DIF statistics are typically removed; items with moderate DIF may be reviewed beyond the normal systematic review process assessing potential item bias, however, with no new information, these investigations typically have low/moderate success. To address this shortcoming, we propose the examination of potential

sources of DIF by using information from the use of Natural Language Processing (NLP) tools to obtain statistical information regarding the underlying causes (e.g., differential knowledge of idioms, linguistic features, word complexity, etc.). NLP tools have never been used to assess sources of DIF in large-scale assessments. We present results from our analyses of a large-scale, high-stakes admissions test with its rising proportion of international examinees. For these examinees, 20%-25% of the items we analyzed had uniform DIF. Non-uniform DIF was detected for an additional 20% of the items. We will present results from the analyses of the sources of DIF, as predicted by the NLP tools. This presentation will advance research on fair assessments by utilizing novel methods to examine sources of DIF for assessments administered to diverse international populations.

Developing Teacher Evaluation Systems: Two New Applications of Standard-Setting

Kimberly Omalley[1] and Kathy Mcknight[2]

[1]Pearson, Austin, United States; [2]Pearson, Fairfax, VA, United States

e-mail: komalley124@gmail.com

Two critical decisions are needed when countries develop teacher evaluation systems: (1) how much weight should be given to each of the multiple measures? and (2) what scores on a measure of teacher effectiveness should be mapped to the policy categories (e.g., effective, highly effective)? These decisions require merging policy definitions, teacher expertise, and empirical evidence. This research adapts research-based methods used to set student performance standards to the application of teacher effectiveness. In this research, we modify research-based methods used to facilitate setting performance standards on student assessments to make the two decisions. The research investigates clear and actionable adaptations to eight common steps in a research-based standard-setting meeting (e.g., briefing book approach). The process requires an iterative approach to (1) defining relative weightings for the various measures used in evaluations, such as value-added model scores, observations, peer evaluations, and surveys and (2) making cut-point decisions on the teacher index to classify teachers into rating categories. We explore optimal ways to incorporate past research on variables related to teacher effectiveness, results from a job analysis, and empirical information, such as the reliability of measures, into the process. Four important changes stood out. (1)This application requires use of different types of empirical evidence. (2) Performance level descriptions span multiple factors related to teacher effectiveness. (3) Pannelists need to provide a broad perspective of teacher effectiveness, and may include as teachers, administrators, students, and policy-makers. (4) Pannelist training should include interpretation of data from multiple sources, understanding limitations of the data, and implications of generalizing the data to the evaluation process as a whole. Applying research-based standard setting procedures to determine weights on multiple measures in teacher effectiveness composites and to map policy definitions of effectiveness categories to composite scores offers advantages over current, policy-driven approaches.

Teaching in a Digital Age: How Educators Use Devices and Implications for Assessment

Kimberly Omalley[1] and Kathy Mcknight[2]

[1]Pearson, Austin, United States; [2]Pearson, Fairfax, VA, United States

e-mail: komalley124@gmail.com

A successful digital conversion for classrooms, districts, and states is not about the technology, it is about how technology enables teaching and learning. Our study took that first step in understanding effective ways to use technology to assess and improve student learning, which is

to identify and define the digital strategies themselves so that they can be measured and connected with student learning. In this study, we conducted focus group interviews to identify digital strategies used by teachers in seven US districts. The forty-four participating educators in these districts have experience teaching in innovative technology classrooms including one-to-one laptops (1:1), bring your own devices (BYOD), blended, flipped, and virtual learning. Through the structured interview process, teachers and administrators described and provided examples of teacher strategies they reported as key factors to positive student outcomes in digital classrooms. We identified seven digital strategies, including using devices to assess, provide rapid feedback, and offer accommodations. Results showed that teachers deploy digital strategies in creative ways, ways that enhance common strategies shown to be best practices based on learning sciences. Findings illustrated that the strategies themselves were not new, the improvements offered by technology were. For example, teachers use chat features, voice-recognition software, and video conferencing to assess and provide rapid, real-time feedback on learning. The seven digital strategies and examples used by some of the most innovative teachers will help new technology adopters understand the wide variety of ways that technology can be used to assess student learning and enhance best teaching practices in a digital age.

Do you Strongly Agree? Gender, Age and Cultural Differences in Response Styles.

Gina Maria Palermo, Tony Li and Alan Bourne

Talent Q Ltd, Oxford, United Kingdom;

e-mail: ginapalermo@talentqgroup.com

Extreme response style (ERS) is the tendency to overuse the endpoints of ordinal response or Likert-type scales. Acquiescence is the tendency to agree on ratings scales regardless of the item content. ERS and acquiescence introduces bias in test takers' scores. If unaccounted for, group variations in ERS and acquiescence can lead to substantial differences in the construct(s) being examined. Thus, the implications of such response styles are key in cross-cultural research. This study investigated the response styles using of a total of 60,786 individuals from seven countries. Gender and age differences in extreme responses and acquiescence response styles were also explored. Marked differences in the tendency to use extreme and 'yea-saying' responses between various nationalities were discovered. No gender differences in ERS were found supporting previous research. However, clear gender differences in both ERS and acquiescence were found across different nationalities. Furthermore, age differences were also shown to have an effect. Implications and recommendations for cross-cultural research in response biases will be discussed.

Effect of Testing on Student Achievement, 1910-2013: Update and Extensions z

Richard Phelps[1] and Monica Silva[2]

[1]Nonpartisan Education Review, Asheville, NC, United States; [2]Pontificia Universidad Católica de Chile, Santiago, Región Metropolitana, Chile

e-mail: richardpphelps@yahoo.com

The paper extends the findings from a recently published meta-analytic review on the effect of testing on student achievement from studies of mastery testing, testing frequency, memory retention, and school and student accountability from quantitative studies, surveys and qualitative studies (Phelps, 2012) (1). The statistical analysis consisted of estimates of effect size based on the results reported in survey and quantitative studies for categories of selected moderators or study artifacts. Mean average and standard deviation for the effect sizes were .56

and .72 respectively. Among the moderators tested at the univariate level, testing with feedback (information and remediation), adding stakes and frequency of testing appeared to strongly and positively affect achievement. The objective of this study was to update and extend the findings of the 175 quantitative research studies to gain a better understanding of the role of moderators in the explanation of variation in effect sizes. The authors will code new moderators and test the joint effects of selected moderators through the application of weighted least-squares meta-regression models using methods outlined by Hedges & Olkin (1985) (2). Preliminary Results In the model built with moderators reported in Phelps (2012) stakes associated to testing is the one with the highest relative weight. The best fitting model accounts for approximately 27% of explained variance. New moderators such as type of design appear to significantly moderate effect sizes. Suggestions are offered for primary researchers to include information on potentially interesting moderators that cannot be tested at present.

Considerations in Setting Standards in International Contexts

Mary Pitoniak[1] and Nan Yeld[2]

[1]Educational Testing Service, Princeton, NJ,United States; [2]University of Cape Town, Rondebosch, South Africa

e-mail: Mpitoniak@ets.org

Standards-based assessments are increasingly being used worldwide to monitor student achievement in relation to set criteria. Cutscores are the points on the score scale that differentiate students into different categories, with often high-stakes decisions being made using these categorizations. Many of the judgmental standard-setting methods used to determine these cutscores have been developed in the United States; however, use of these methods may be complicated by factors related to the political and educational contexts found in different countries. Consideration should be given to how to modify these methods as needed in a given international setting. This paper describes standard setting activities undertaken for criterion-referenced tests developed in South Africa to help post-secondary institutions meet the educational needs of their incoming students. Within South Africa, the educational system reflects the changing social and cultural conditions of the post-apartheid years. Inequities are being addressed, but there are challenges, and realities that must be faced. This changing context amplified issues that often arise in U.S. standard setting implementations. Issues that are discussed in the paper include (1) defining borderline performance within a very diverse educational system, (2) acknowledging Pannelist anxiety and discomfort when making judgments about students, particularly those that place students into the lowest performance levels, and (3) assembling a Pannel that reflects the social diversity of South African universities. The educational context of a given country has a large impact on all aspects of large-scale assessment. While issues unique to the South African context amplified difficulties during various steps of the process, the results can provide useful reminders to anyone conducting a standard-setting study. Heeding these reminders will enhance the validity of the interpretations made on the basis of the classifications into the performance levels.

The Underlying Latent Structure of Tests for Non-Cognitve Constructs in Education

Ricardo Primi[1] and Daniel Santos[2]

[1]Universidade São Francisco, Itatiba, Brazil; [2]Universidade de São Paulo, Ribeirão Preto, Brazil

e-mail: rprimi@mac.com

Policymakers have been increasingly interested in incorporating socio-emotional measurements in assessment systems to monitor educational organizations as well as broad level social outcomes. A recent movement called partnership for the 21st century skills has emphasized the need for the development of large-scale socio and emotional instruments for assessing children in their schools. Literature reviews proposes the understanding of these skills within the framework five-factor model. This study investigates the factor structure of constructs used in educational settings - self-concept, self-efficacy, self-esteem, motivation, and attitudes and beliefs control of events, adaptability and the big five - to test whether the big five dimensions framework would be able to explain the underlying structure among the subscales. The sample was N=3023 children from five grades (5th N=697, 6th N=710, 9th N=674, 10th N=488, 12th N=454), from 16 Schools and 86 classes. They answered to the Big Five Inventory (BFI), Big Five for Children (BFC), Strengths and Difficulties Questionnaire (SDQ), Self-Efficacy Questionnaire for Children, Grit Scale, Core Self Evaluations, Rosenberg Self-Esteem Scale, Norwick-Strickland Locus of Control Scale (total number of items n= 209). A Balanced Incomplete Block (BIB) design was used to systematically produce booklets of tests that contained a median of 67 items while preserving all the combinations of each two tests making it possible to calculate the complete inter-test correlations matrix while not exhausting the children with too many items to answer. Parallel analysis justified the five-factor solution. An Exploratory Factor Analyses (EFA) of this solution could clearly interpreted according to the five-factor model: Conscientiousness and openness is linked to the ability to self-motivate; extraversion to the ability to initiate social interactions; neuroticism to the ability to deal with negative emotions and beliefs of control of events; agreeableness to the ability to maintain positive social skills and interactions.

Application of the Logistic Regression Procedure for Assessing Differential Item Functioning in Computer Adaptive Tests

Joseph Rios

University of Massachusetts Amherst, MA, United States

e-mail: jarios@educ.umass.edu

A dearth of research has been conducted in the area of assessing DIF in a CAT environment (Gierl, Lai, & Li, 2013). The few procedures that have been suggested suffer from both practical and methodological constraints. For example, many of the procedures require data imputation to match examinees across a complete item bank, while others cannot be fully implemented in operational contexts due to software limitations. As a result, further research is needed in this area (Zwick, 2010). This study investigated the adequacy of applying the logistic regression DIF procedure for evaluating pilot items in a CAT (CAT-LR). As examinees cannot be matched on total score in a CAT context, the CAT-LR procedure matches examinees on a theta estimate obtained from a CAT administration of operational items. The adequacy of this procedure was evaluated for type I error, power, correct detection of DIF type, and effect size classification accuracy rates using Jodoin and Gierl's (2001) guidelines under four independent variables: 1) sample size, 2) degree of impact, 3) DIF item type, and 4) DIF magnitude. Inflated type I error (25% to 33% on average) and decreased power rates (as low as 72%) were exhibited under impact. Furthermore, low true-positive rates (5% to 7%) were exhibited for accurately identifying uniform DIF items. Lastly, under impact, DIF magnitude was underestimated at a rate around 100%. Results from

this study suggest that the CAT-LR procedure was inaccurate in correctly identifying DIF, DIF item type, and DIF magnitude when impact was present. These results suggest that this method may be inappropriate for: 1) purifying items that possess construct-irrelevant variance and 2) relating DIF results to test development and substantive research. Future research directions in this area will be addressed in the full paper.

Do Competencies Predict Objective Performance of Employees?

Yasin Rofcanin, Levent Sevinç and Aykut Berber
Assessment Systems Turkey, Istanbul, Turkey;
e-mail: yrofanin@yahoo.com

To date, studies on competency models have focused on the development of general models (e.g., Redmond, 2013) which predominantly focused on aligning key competencies to strategies, goals and performance of companies (e.g., Becker & Huselid, 1999; Campion et al., 2011). This research aims at contributing to this gap by validating a new measure on sales competencies (SALECOM) developed across industries in Turkey. In particular, our goal is to contribute to our understanding on sales competencies via establishing links between sales competencies and outcomes. We carried out two studies to realize our research goals. In Study 1 (N = 336), we examined the factorial structure of SALECOM and established its link with objective performance outcomes which were measured as percentage of sales goals realized. Moreover, in this study, we controlled for the effects of self-promotion and self-modesty. In Study 2 (N = 527), our purpose was to strengthen the validity of SALECOM by linking its explanatory power to performance outcomes different from Study 1. Findings Regarding Study 1, findings from confirmatory factor analyses (CFA) revealed a satisfactory fit of SALECOM to data, with all items loaded onto their pre-defined dimensions (χ^2 = 414, df = 131, GFI = .95, SRMR = .05, RMSEA = .05, CFI = .96). Moreover, SALECOM explained 8.5 % variance in performance (R^2=.085, $p<.001$). In Study 2 (N = 527), we attempted to replicate our previous findings by including a broader range of objective performance outcomes (i.e., percent of net income over budget; sales per number of invoices; number of items sold per number of invoice). With this study, we make at least two contributions. Our contribution relates to the development of a new competency model regarding sales functions. This is critical because previous research has focused on general competency measures.

Improving the Utility of Large-Scale Assessments

Todd Rogers
University of Alberta, Edmonton, Alberta, Canada
e-mail: todd.rogers@ualberta.ca

This proposed presenatation is a position paper. Principals and teachers do not use large-scale assessment results to the extent desired because (a) the lack of distinct and reliable subtests prevents identifying strengths and weaknesses, (b) the results arrive too late to be used, and (c) they need assistance to use the results to improve curriculum and instruction so as to improve student learning. Most assessment officials do not deliberately determine if the assessment domain is multidimensional. Procedures used to determine if the subtests are distinct (e.g., agreement method (Kelly, 1923; Gulliksen, 1950; Lord & Novick, 1968); correlations corrected for attenuation (Haladyna & Kramer, 2004); and generalizabilty analysis (Rogers & Radwan, 2012)) reveal that subtests are not distinct nor do they add value over the total score (proportional reduction in mean square error (Sinhary, 2010)). Principals and teachers tend not to use the results to make changes in what and how they teach (Klinger & Rogers, 2011). It is

recommended that (a) the first activity of an assessment be to clearly establish that the domain is clearly multidimensional, (b) the assessment schedule be changed so that a given subject area is assessed in non-consecutive years but the number of sittings each year remains the same, and (c) assistance be provided to principals and teachers to allow them to use the assessment results to improve curriculum and/or instruction so as to enhance student learning and achievement. Three suggested assessment schedules are presented in which three or four assessments are administered with at least two years between consecutive assessments of the same subject area. Implementation of any of the schedules would increase the reliability of subtests to at least 0.80 and allow principals and teachers at least two years to plan and implement changes in curriculum and instruction.

Impact on Equating Results by Excluding Repeaters and Students with Specialized Accommodations from the Equating Sample

Todd Rogers[1] and Nizam Radwan[2]

[1]University of Alberta, Edmonton, Alberta, Canada; [2]Education Quality and Accountability Office, Toronto, Ontario, Canada

e-mail: todd.rogers@ualberta.ca

Restricted equating samples are often used to equate tests. Previously eligible students are excluded because this group of students is not stable from year to year. Students receiving specialized accommodations may be excluded because of the nature of their accommodations. The objective was to evaluate the impact of including previously eligible and students with specialized accommodations (first time studied) in equating samples. Target populations consisted of first time eligible students, previously eligible students from 2011, and students with specialized accommodations who need to write the Ontario Secondary School Literacy Test/Test provincial de competences linguistiques (OSSLT/TPCL) administered at Grade 10 in order to graduate. Four equating samples were considered: first time eligible and first time eligible with previously eligible, students with specialized accommodations, and both previously eligible and students with specialized accommodations. The forward-fixed common-item parameter procedure for non-equivalent groups was used to equate the 2012 "operational" form with the 2011 operational form, where the items in the 2012 form were embedded field test items in the 2011 forms. Root mean square difference, conditional standard error equating, and percent passing were used to compare the four sets of results. Including previously eligible students and/or students with specialized accommodations in equating samples had little impact (e.g., difference between theta values between pairs of samples are no greater than 0.10 for vast majority of students; percentages of successful students in the population are within 0.2% of each other for the four OSSLT samples and 0.3% for the four TPCL samples). It is reasonable to include previously eligible and students with specialized accommodations in the equating samples, thereby increasing the representativeness of the equating samples. Given the percentages of these students were not large (< 20% when included together), the findings may not be hold where percentages of excluded students are larger.

You May be Able to Convert Your Fixed Form test to a Computer Adaptive Test

Lawrence Rudner

Graduate Management Admission Council, Reston, VA, United States

e-mail: LRudner@gmac.com

Many testing programs that use fixed form tests do so because they believe they do not have enough items or enough test takers to administer a computer adaptive test. If Item Response

Theory is the underlying model, then indeed 1000 to 2000 test takers per item and a bank of several hundred items are often needed. But IRT is not the only model available for CAT. If the goal is simple classification, e.g. pass/ fail, or A/B/C/D, then a simpler model will often work. This paper will present and evaluate the use of Measurement Decision Theory as the underlying model for a CAT program. With MDT, item parameters are the group-conditional probabilities of a correct response rather than IRT probabilities across a continuum. Ability is estimated in terms of the most likely classification rather than a point estimate along a continuum. Relative to IRT, MDT should require fewer examinees for calibration and fewer items for classification. Data from an intact certification program is used to illustrate and evaluate MDT CAT. Items were calibrated as the probabilities of a correct response for those that were certified in the past and the probabilities of a correct response for those that failed the examination. Items were selected adaptively using Kullback-Leibler to yield the highest expected post-administration information. To illustrate the power of this methodology, the number of items needed to be administered in order to achieve 90 percent classification accuracy and the number of examinees needed to achieve satisfactory conditional probabilities are examined. The MDT results are compared with the current test length and current calibrations sizes used by this program. Initial analysis shows that a 75% savings in test length and sample size could be achieved with this model.

Are the Professionals Ready to Use our New Technologies of Assessment? Survey of Spanish Psychologists About its Use of New Technologies

Pablo Santamaría[1] and Rodolfo Ramos[2]

[1]TEA Ediciones, Madrid, Spain; [2]Universidad de Granada, Melilla, Spain

e-mail: pablo.santamaria@teaediciones.com

Over the past decade, major advances have occurred in the development of new technologies in the area of psychological assessment. Nevertheless, very little systematic research has been carried out to study the level of readiness of the professionals to use new technologies and, particularly, in the assessment area, despite the practical relevance of the topic. This research was aimed to obtain preliminary information about the use of new technologies by Spanish psychologists. A 66-item questionnaire was created to enquire about the general level of use of new technologies (Dropbox, Skype, IOS, Android, Moodle...) and of assessment and intervention new technologies (on-line assessment, on-line scoring...). All the items were Likert-type with five categories, scored from 0 to 5 according to their level of use of that technology (0=none ; 1 = very low ; 2 =low; 3=average; 4=high; 5=very high). A small tryout study (n=46) was carried out to ensure the survey items could be perfectly understood and interpreted by the population they addressed. The sample was made up of 1,728 professional psychologists (73% women; 27% men) who responded to a survey sent to the members of the Spanish Psychological Association (COP). The mean age was 42.8 years and the standard deviation 10.7. As regards the field of specialization, 56% work in Clinical Psychology, 12% in Educational Psychology, 11% in Social Services, 8% in Work and Organizational Psychology, and 13% in other fields, such as sports, forensic or traffic. Descriptive statistical analyses were carried out on the items and on the general data requested from participants. General results showed a moderate or low use of new technologies by Spanish Psychologist (e.g. Dropbox M=2.0, Sd=1.9; on-line assessment M=0.9, Sd=1,5; on-line scoring M=2.1, Sd=1.9) that will be discussed in detail.

Situational Judgment Tests: the Pros and Cons of a Construct-Driven Approach

Lolle Schakel and Annette Maij-De Meij
PiCompany, Utrecht, Netherlands
e-mail: l.schakel@picompany.nl

Situational Judgment Tests (SJTs) are a popular and useful method for identifying qualified applicants. They present applicants with work-related situations and ask them how they would respond. Despite the widespread use of SJTs, test developers and researchers often give little attention to the constructs measured by SJTs and tend to report results based on overall scores. However, to understand how and why SJTs work in selection context, it is critical to identify the constructs assessed by SJTs related to the performance domain. This study is innovative in that a construct-driven approach is used to develop a video-based SJT for communication styles for the selection of call center employees. The construct validity of the SJT is investigated using data of 147 non-call center employees who completed the SJT and the Communication Styles Inventory (CSI). Contrary to our expectation, results showed no significant correlations between the scores on the domain-level scales of the SJT and the scores on the corresponding scales of the CSI. Also, for a sample of 146 call center employees no significant correlations were found between the overall SJT score and job performance. However, call center employees scored significantly higher on the SJT, which indicates that the SJT is able to distinguish between experienced and non-experienced employees. A possible explanation of these results may be related to differences in response instructions of the SJT and the CSI (level of agreement). In addition, the call center context of the SJT might elicit more socially desirable behavior than the context-free statements in the CSI. Summarized, the SJT might have resulted in the measurement of behavioral effectiveness instead of communication styles. Overall, these findings suggests that SJTs could be more appropriate to measure constructs like personality or job knowledge, than measuring other constructs like communication.

Evaluating the Utility and Scoring of Technology Enhanced Items for Computer-Based Testing

Jason Schwartz[1], Hong Qian[2] and Joe Betts[1]
[1]Pearson VUE, Chicago, United States; [2]National Council of State Boards of Nursing, Chicago, United States
e-mail: jsmathematics@gmail.com

As assessments move from traditional paper-and-pencil formats to computer-based testing (CBT), the opportunity to utilize the unique characteristics of the computer to deliver technology enhanced items (TEI) emerges. However, little research has focused on the possible implications of using these expanded delivery options. This presentation will highlight some of the emerging options for TEI within the context of CBT and discuss possible scoring options for these newer item types. The presentation will discuss the structure of a number of new item types that could be easily implemented in CBT and discuss their utility for different assessment goals, e.g. educational and psychological tests focused on individual differences or licensure/certification exams that focus on determining minimum levels of competence to make decisions about candidate proficiency for entry-level practice. Each item type will be prototyped to provide a visual instantiation of the structure of the item that could be realized in an operational setting. This will provide participants a context to evaluate the elements of each item type. In addition, a discussion of a number of possible approaches to scoring those item types will be undertaken. A critique of each method of scoring will be presented along with practical considerations for each method within the context of different testing goals. Participants will be exposed to a number of

new technology enhanced items along with a number of options for scoring those different item types. Practical discussions will ensue to provide a framework for evaluating the possible best case uses of the different item types and evaluate the tradeoffs for the different scoring options.

Rater Effects in High Stakes Post-Secondary Admissions Testing: "Finding Beauty in the Beast"

Stefanie Sebok, Stefan Merchant and Don Klinger
Queen's University, Kingston, Ontario, Canada
e-mail: stefanie.sebok@queensu.ca

Many post-secondary institutions use some form of admissions "testing," especially for specialty and high demand programs. These tests include writing samples or other measures of performance assessment. Trained raters complete the evaluation of these assessments. Assessment and testing situations that require the use of raters are susceptible to issues of validity and reliability given that raters are highly variable in their interpretations. Although efforts have been made to reduce unwanted rater variability in high-stakes situations, many of these attempts (e.g., rater training) have demonstrated minimal success. One of the areas that has been identified, yet underexplored, is the cherished value systems of raters. This study investigates how raters' personal values and beliefs are reflected in their scoring of writing samples and how training impacts the operationalization of those value systems during the rating process. To better understand how raters perceive and categorize the writing samples of applicants, a pile sorting method was employed. Each participant was given a set of writing samples and asked to divide them into a number of categories that best represent different types of applicants. Participants were permitted to have as many categories as they wish and were asked to explain their categorization criteria. Using the same set of writing samples, the participant was asked to divide them into three categories: accept, do not accept, and undecided as well as provide a reason for the category assignment. Although analysis of the data is ongoing, through the use of cluster analysis consensus among the participants' categorization techniques has been identified. This study has important implications for explaining how raters translate their values, beliefs, and instincts as they formulate a judgment about an applicant. Further, this study could be used to inform rater training and other efforts to reduce undesired rater variability.

Predictive Power of AC over Supervisor Rated Performance: A Multi-Source and Multi-Method Study in Turkey

Levent Sevinç and Yasin Rofcanin
Assessment Systems Turkey, Istanbul, Turkey
e-mail: levents@assessment.com.tr

Rise of uncertainity, fierce competition and dynamism pressure employees to drive their performance (e.g., Pulakos, Arad, Donovan, & Plamondon, 2000). As performance has become a critical evaluation tool for employers, assessment centers (AC) have been increasing utilized to evaluate employee performance on many criteria (Povah & Thornton, 2011; Griffin, Neal, & Parker, 2007). Yet, especially in emerging country contexts, research is insufficient with respect to the link between ACs and performance outcomes that are central to organizations. Accordingly, aims of this study were two-fold. Our first aim was to examine the predictive power of AC over supervisor-rated performance. Our second goal was to see if participants who obtained low and high scores from ACs clustered in separate groups. Data were drawn from a single company based in Istanbul, Turkey (N = 97) within an AC setting. In a separate time, we

obtained objective performance scores of participants from their immediate managers. Findings from multiple regression analyses demonstrated that AC accounted for a significant variance in performance (Adjusted R2 = .27). Our findings revealed that ACs predict key objective performance. Moreover, there is a great degree of overlap between the predictive power of AC results and two clusters emerged. Majority of the high performers were clustered in the first group who also obtained high competency scores from in AC.

Measuring Students' Adjustment Strategies and Number Sense by Online Estimation Assessment with Broken Calculator Problems

Shu-Chuan Shih, Shu-Juan Lee and Bor-Chen Kuo

Graduate Institute of Educational Measurement and Statistics, National Taichung University of Education, Taichung City, Taiwan

e-mail: ssc@mail.ntcu.edu.tw

Number sense is considered to be a significant topic in mathematics education recently (National Council of Teachers of Mathematics, 2000). This term has been described by McIntosh, Reys, Reys, Bana and Farrell (1997) as "a person's general understanding of number and operations, along with the ability and inclination to use this understanding in flexible ways to make mathematical judgments and to develop useful and efficient strategies for managing numerical situations". Based on this definition, computational estimation has been considered as an important indicator to assess number sense of students (Howden, 1989). Therefore, the purposes of this study are to construct an online estimation assessment with broken calculator problems that can elicit students to use adjustment strategies in estimation process demonstrating their number sense, and design the autoanalysis mechanism for data gathered from students' problem-solving log files. There are 58 computational estimation problems constructed to explore adjustment strategies used by students, and the assessment is administered to 724 third and fourth grade students in Taiwan. The Cronbach's value of the assessment is .91. This indicates good internal consistency for items in this assessment. Students' number sense is measured using the computerized number sense multiple assessment system proposed in the previous study (shih, 2013). The measurement model is one-factor hierarchical item response theory. The results are shown as follows: 1.The adjustment strategies performances in estimation and number sense of fourth graders are significantly better than those of third graders in Taiwan. 2.The distributions of various adjustment strategies in estimation used by examinees with different genders and different number sense (high-ability group vs low-ability group) are also discussed to afford informative diagnosis.

Validating the Objective Borderline Method (OBM) for Standard Setting Using a Simulation Study

Boaz Shulruf[1], Phil Jones[2], Phillippa Poole[3] and Tim Wilkinson[4]

[1]University of New South Wales, Sydney, Australia; [2]UNSW Medicine, Sydney, Australia; [3]University of Auckland, Auckland, New Zealand; [4]University of Otago, Christchurch, New Zealand

e-mail: b.shulruf@unsw.edu.au

Setting standards in educational assessment is one of the most challenging branches of psychometrics. The main challenge is that standard setting intertwines subjective judgement with statistical procedures in a decision making process which aims to provide defensible, valid and reliable outcomes. Research demonstrates that whenever two or more different standard setting methods are applied to the same dataset, each produces a different cutscore, which questions the validity of such methods. Recently a probabilistic standard setting method, the

Objective Borderline Method (OBM), was introduced and presented as having advantages over more commonly used methods (e.g. Angoff, Regression etc). The current study utilises simulated data to test the validity of a modified version of the OBM (mOBM). This involves 2500 different datasets each comprised of 1000 simulated students' examination scores. The simulation parameters determined student ability, a borderline range (per examination) and examination scores. The validity measure compared student 'true' ability scores with the Pass/Fail decisions for Borderline scores derived from the mOBM. The result demonstrated that, on average, the mOBM correctly classified 70% of the Borderline grades. 3.4% of the Borderline scores were reclassified as Pass where 'true' ability was at a Fail level. On the other hand 26% of the Borderline scores were reclassified as Fail where 'true' ability was at a Pass level. Across all datasets, the mOBM achieved specificity of .88 and sensitivity of .51. The mOBM was not affected by the size of the borderline range. This simulation shows that the mOBM has validity as a standard setting method. Currently it is not possible to compare its validity to other standard setting methods as no similar validity studies on simulated data (comparing 'true' ability with Pass/Fail decisions) were found in the literature. Further research using simulated data for validating and comparing standard setting methods is needed.

Evaluating Consequential Tests and Conflicts of Interest: The Case of Chile's PSU

Monica Silva[1], Richard Phelps[2] and Mladen Koljatic[1]

[1]P. Universidad Catolica de Chile, Santiago, Chile; [2]The Association of Boarding Schools, Asheville, NC, United States

e-mail: msilvara@uc.cl

Countries differ in the degree to which they can exercise statutory control over the use of testing and its consequences for those tested (ITC, 2013). In some nations, such as Chile, standards to guide test development and administration do not exist and might not be legally enforceable even if they did. In that context, independent evaluations of high-stakes tests may be of paramount importance to guarantee the rights of test-takers to be assessed with reasonably valid and fair tests. The Chilean Admission Test (PSU) is a controversial curriculum-based high-stakes test required for admission universities. Between 2004 and 2013 eight evaluative studies of the PSU were conducted: two performed by international agencies (ETS in 2005, Pearson in 2013). In the case of the ETS audit, the assessors were paid by the test developer; in the case of Pearson the request for proposals were written by the test developers and administrators. The authors explore the consistency (and inconsistencies) among the official reports and the independent audits. The analysis focuses on their overall appraisal of the quality of the PSU as an admission test and their assessment of evidence regarding validity and fairness issues. The case study is primarily construed on the bases of secondary analysis from archival data of evaluations, technical reports and press releases. The analysis indicates wide inconsistencies between official evaluations and the international audits in their assessments of the PSU's relative quality, validity, and fairness. But, even the international audits lack some key policy-relevant information and do not prioritize the changes called for. Governments should invest in periodic independent evaluations of consequential testing programs. Resources for the evaluations should be allotted from program inception and measures taken to ensure that those assessments are genuinely independent and free of conflict of interest.

Good Practice: Using ISO 10667 to Implement Mechanical Data Combination in Assessment Center

Anders Sjöberg[1] and Eva Bergvall [2]

[1]Stocholm university, Stockholm, Sweden; [2]CLU, Göteborg, Sweden

e-mail: anders.sjoberg@psychology.su.se

The international standard ISO 10667: "Assessment service delivery: procedures and methods to assess people in work and organizational settings", cover procedures and methods in assessment within workplace settings. Assessment Center (AC) is one example of a process description of a method which is included in ISO 10667. AC is designed to measure multiple dimensions (e.g., problem solving and interpersonal skills) through exercises for prediction of future performance on the job. Although AC dimensions shows an incremental validity over and above psychometric tests measuring cognitive ability and personality, the actual decision making is often based upon consensus discussions among raters (Dilchert & Ones, 2009).This approach, referred to as clinical combination of data in the research literature(e. g., Meehl, 1954; Sawyer, 1966) is still the predominating approach in practice, AC is no exception. For prediction purposes however, an opposed mechanical approach using a mathematical algorithm to combine data, has proven superior to clinical data combination for a wide range of criteria (Sarbin, 1943; Gough, 1962; Meehl, 1965, 1967; Sawyer, 1966; Goldberg, 1968; Sines, 1970; Dawes et al.,1989; Grove & Meehl, 1996;; Grove, Zald, Lebow, Snitz, & Nelson, 2000) including job performance (Kuncel, Klieger, Connelly, & Ones, 2013)Based on research and the formulation in the ISO standard about decision making, this study illustrates how to implement an Evidence based AC (EAC), using mechanically combined AC ratings to predict managerial performance.

The Susceptibility of Performance Items to Exposure

Russell Smith

Alpine Testing Solutions, Henderson, NV, United States

e-mail: russell.smith@alpinetesting.com

Regardless of the specific terminology utilized to characterize unintended advantages - exposure, prior knowledge, cheating, item parameter drift - security concerns in information technology (IT) certification programs are pervasive. Items, forms, and even item banks are often available on the Internet within days of exam publication. As one way of minimizing security risks, Cizek (1999) suggests using "a variety of other assessment approaches" and notes that "performance assessments require students to actually demonstrate their knowledge or skill" (p. 168). Given the content domains and computer-based delivery mode, IT certification exams lend themselves to the development and administration of performance item types, often simulations or emulations. However, little is known about the susceptibility of these item types to exposure. Hypotheses exist that these items are more memorable, and therefore more susceptible to exposure. Others argue that these items are less at risk as even if a candidate has prior knowledge, they still need to complete the task. This study will explore the susceptibility of performance items to exposure as compared to select-type items. This is an extension of prior research which explored the difference between select-type items and simulation items for one exam. Figure 1 shows the moving average score averaged within item type. The figure demonstrates the two main findings about simulation items from that study: 1) they tend to be more difficult and 2) their difficulty tends to be more stable than select-type items. The proposed study will attempt to generalize these findings across multiple exams and programs as well as extend this research to include additional analyses, such as IRT drift analysis.

An Adaptive Item Selection Method for Curtailment*

Niels Smits
VU University Amsterdam, Amsterdam, Netherlands
e-mail: n.smits@vu.nl

Health questionnaires are often built up from sets of questions which are totaled to obtain a sum score; often, this score is subsequently used to classify respondents. An important consideration in designing questionnaires is to minimize respondent burden. Finkelman et al. (2012) introduced curtailment as an efficient method of questionnaire administration aimed at classified as ' at risk' and ' not at risk'. Curtailment uses a prediction model for forecasting observed class membership; the strategy is to stop testing when not yet administered items are unlikely to change the respondent's classification of curtailment is static, i.e., is equal for all respondents, and dynamic item selection could increase efficiency. The current paper uses a method for adaptive item selection which stems from Data Mining (Hastie et al., 2009). The item selection method will be studied using several real data sets. Benefits and limitations of this methodology are discussed.

Do Cognitive Processes Involved in Solving Reading Comprehension Items Differ in Students with Differing Language Background?

Philipp Sonnleitner, Gina Wrobel and Monique Reichert
University of Luxembourg, Luxembourg
e-mail: philipp.sonnleitner@uni.lu

One major global challenge of educational assessment that has to be addressed on a local level is the increasing number of students with immigration background usually speaking a different language at home compared to their native peers (OECD, 2012). Especially in large-scale contexts, however, individual and tailored testing responding to their specific (language) needs is not possible. Although DIF-analyses are common practice in current large-scale assessments, they only indicate whether and to what extent an item is biased but provide no information on which cognitive processes might cause that bias – crucial information when evaluating school systems. The current study goes beyond traditional DIF-analyses by using the IRT based linear logistic test model (LLTM; Fischer, 1973) that allows for modeling cognitive demands and therefore processes involved in each item. Specifically, we draw on a sample of more than 5000 Luxembourgish 3rd graders and analyze whether cognitive and linguistic item attributes (e.g., kind of inference that is needed to solve the item, textual coherence; Sonnleitner, 2008) of a large-scale reading comprehension test do possess different difficulty for students with varying language background. We do this by determining a cognitive model including such attributes that adequately describe item difficulty parameters in native students. Subsequently, we will cross validate this model in several sub-samples with varying language background. Results not only show if cognitive and linguistic item attributes do differ with regard to difficulty in the different samples but also if some cognitive processes do compensate each other in certain samples. It will be discussed how these results can be used to complement common DIF-analyses and to obtain more fine-grained information on students' performance differences in reading comprehension.

Examining Effects of Formative Feedback with Item Response Trees

Claire Stevenson[1] and Paul De Boeck[2]
[1]Leiden University, Leiden, Netherlands; [2] The Ohio State University, Columbus, United States
e-mail: cstevenson@fsw.leidenuniv.nl

Computer-based formative assessment utilizes various forms of feedback and has enormous potential in optimizing learning by providing feedback tailored to an individual's instructional needs. However, determining what type of feedback best optimizes the learning of a particular task for a particular individual is a complex endeavor. The effectiveness of different types of feedback is not always clear-cut. Furthermore, individual differences may be present in how effective each of these types of feedback is at different stages in the learning process. This is further complicated by the need to account for the effect of item characteristics on feedback effectiveness. Explanatory item response trees allow for the inclusion of both learner and item characteristics while examining learning as it occurs during computerized training. In the present study we used explanatory item response trees to model children's change in analogical reasoning on a trial-by-trial basis during the course of training. 705 children (M=7.5; range 5.0-11.0 years) solved 10 analogy items and received adaptive feedback per trial (maximum 5 trials) in the form of graduated prompts (N=505) or outcome feedback (N=200). The utilized item response tree models take revised responses after feedback into account and examine the effectiveness of different types of feedback based on learner and item characteristics. Results show that metacognitive prompting and outcome feedback were similarly effective in inducing correct solutions. More specific cognitive prompts and scaffolds led to correct solutions more often than outcome feedback. Furthermore, the effectiveness of cognitive prompts and scaffolds appears related to initial ability, cultural background, working memory efficiency and age. Item response trees appear an effective tool to analyze feedback effects during formative assessment.

Challenges to Best Practice in Assessment: Act Locally--Think Globally

Donna Sundre and Sara Finney
James Madison University, Harrisonburg, VA, United States
e-mail: Sundredl@jmu.edu

While ITC conference attendees recognize the psychometric rigor required for sound assessment practice, review of actual practice produces highly disappointing results. The purpose of this paper is to explicate our use of the scientific method to obtain high quality assessment data. We have employed these methodologies for over 25 years at the largest center for higher education assessment in the United States, and possibly the world. Our contention is that there are three prerequisites for credible assessment data: 1) sound sampling from the target population; 2) psychometric excellence—including demonstrated linkage to student learning domain and learning opportunities; and 3) examinees who are motivated to perform well. All three are required. The paper/presentation will present our standards for achieving best testing practice. The populations we wish to make inferences about are: 1) entering first-year university students and 2) second-year students; these form pre-test/post-test groups for assessing our student learning objectives. Students are required to participate in this assessment, but no student will take all assessment tests. Entering students are assigned to testing rooms via the last 2 digits of their student IDs-resulting in large, representative samples; these same students are re-assessed at the end of their second year using the same instruments. Over 90% of the instruments employed have been developed by university faculty working with measurement experts. All students also complete a measure of examinee motivation. By coupling various methodological approaches, we have garnered evidence to support student learning during the

first two years of college. That is, our research highlights how the examination of test scores using different approaches facilitates valid inferences from the scores. The key to achieving assessment best practice is to be guided by a commitment to quality through use of the scientific method to obtain credible high quality data.

Item Format Effect over the Social Desirability Assessment

Álvaro Villegas, Álvaro Postigo, Javier Suárez-Álvarez and Eduardo García-Cueto
Universidad de Oviedo, Oviedo, Spain
e-mail: alvarovillegasf@gmail.com

The problem of social desirability is, and has been, constant in the field of psychological assessment. It is well known that, in one way or another, it could be affecting any kind of test; however it is clearly underscored to a greater extent when the results of the assessment entail direct and important consequences on people's future. This is the case, among others, of the personnel selection process in which personality treats are kept in mind. Numerous methods have been attempted in order to try to detect and minimize the effect on social desirability in tests measurement. The aim of this current research is to try and determine to what extent detecting social desirability depends on the method used and, if so, which would be the most appropriate. In order to achieve the previously exposed aim, a social desirability scale was developed, which was applied following three different normative, ipsative and the Thurstone binary comparisons. The obtained results show that the ipsative format is the one which seems to produce the worst results when detecting social desirability on contestants of a test. In spite of the defence promulgated on the ipsative formats in order to reduce social desirability, in the case of the lie scales it seems, this, to be the less efficient to reach that purpose.

Psychometrics for a New Generation of Assessments

Alina von Davier
Educational Testing Service, Princeton, NJ, United States
e-mail: avondavier@ets.org

As education enters a new world defined by technology driven innovations, test publishers need to understand how to synthesize advances in technology, statistics, and the learning sciences to support the creation of a new generation of assessments. I will consider some of the specific features of the new types of assessments. These include: • Psychometric and statistical models appropriate for process data, such as data collected from simulations, educational games and intelligent tutoring systems • Data mining techniques and dynamic models for 'big data' Examples of current projects that address these features will be mentioned. The presentation will close with a more detailed examination of one of the challenges for the new assessments — measuring individual cognitive skills through collaborative problem solving (CPS) tasks. It is generally recognized that working well with others is a key 21st-century skill, but what are the appropriate psychometric models for assessments of cognitive skills as inferred from CPS? This research focus is exemplified by a new project underway at ETS, the Tetralogue. The Tetralogue is a science assessment package that consists of traditional assessment components and a simulated science task that allows for both human to human and human to computer-agent interactions. The Tetralogue also includes a non-cognitive (personality) measurement instrument. This project aims to fill the gaps mentioned above. Preliminary results of our research will be presented.

Testing for Bandwidth and Fidelity in Personality Inventories: A Bifactor Model

Paul Vorster and Gideon P. De Bruin
University of Johannesburg, Johannesburg, South Africa
e-mail: paul@jvrafrica.co.za

Personality instruments are believed to measure at a high or low level of measurement abstraction, but not at both levels simultaneously (Cronbach & Gleser, 1965). This influences the degree of generality or specificity of personality instruments which is referred to as the bandwidth/fidelity dilemma (Judge et al., 2013). This dilemma is especially important for hierarchical personality models as they claim to possess both broad bandwidth and high fidelity. Unfortunately, this implicit assumption has not been investigated on a construct level. (Paunonen & Ashton, 2013). This study aims to ascertain whether a hierarchical personality test, the Basic Traits Inventory (BTI), incorporates both bandwidth and fidelity at the higher and lower personality measurement levels. For hierarchical personality tests to have both broad bandwidth and high fidelity the general factors should not explain all of the common variance but should allow sub-factors to explain some of the common variance above and beyond the general factors while maintaining unidimensionality at the factor level. Data were sourced from a psychometrics database of South African adults (n = 1962). The data were subjected to three confirmatory factor analytic models, namely (a) a one-factor model, (b) a multi-factor model; and (c) a restricted confirmatory bifactor model (cf. Reise et al, 2010). The bifactor model fit the data best and the BTI scales maintained their unidimensionality at the factor level, but also allowed the sub-factors to explain common variance exclusively. This indicates that the BTI may measure with broad bandwidth at the factor level and high fidelity at the facet level. The use of the confirmatory bifactor analytic model can be used to ascertain whether hierarchical personality tests measure broadly at the factor or scale level and narrowly at the sub-factor or facet level. Implications for analysis of tests for bandwidth and fidelity are discussed.

Comparison of Item-level and Multistage Computerized Adaptive Testing with Complex Test Constraints

Ada Woo, Xiao Luo and Philip Dickison
National Council of State Boards of Nursing (NCSBN), Chicago, United States
e-mail: adawoo@outlook.com

The advent of inexpensive computing devices and continuous developments in information technology have resulted in a variety of innovative test delivery modes, providing more efficient measurements than the traditional paper-and-pencil test. Among all, the computerized adaptive testing (CAT) is well acknowledged as a very promising replacement of the traditional test in which each test form is adaptively assembled in real time to match the test-taker's ability. Since the customization of test form might introduce biases, incomparability, and unfairness to the test score and pertinent score-based decisions, various content-based and statistical constraints are typically incorporated into the item selection rules in order to ensure the equality of different test forms. The purpose of this study is to investigate the effect of the complexity of test constraints on two types of CAT, namely item-level and multistage CAT. The complexity of constraints is manipulated to vary from a simple content-balancing constraint to multiple complex constraints regarding content balancing, response time, enemy item, item set, etc. A realistic operational item pool of a large-scale high-stakes licensure examination is used as the item pool. In each condition, 10,000 examinees are generated and tested with two simulators that are programmed in R to simulate the item-level and multistage CATs respectively. Results are evaluated in terms

of ability estimation, constraint violation, item exposure, item usage rate, etc. We expect that the item-level CAT would deliver better overall performance when the constraints are simplistic and manageable in relative to the resources available in the item pool. As constraints become more complex and the item selection becomes more restricted, multistage CAT would deal with the complex item selection more efficiently with a global optimization algorithm than an item-level CAT. Practitioners should choose the appropriate CAT after considering the requirements and resources of testing programs.

Latent Class Structural Equation Modeling as a Tool for Developing Validity Arguments

Amery Wu[1], Jake Stone[2] and Yan Liu[3]

[1]The University of British Columbia, Vancouver, Canada; [2]Paragon Testing, Burnaby, Canada; [3]UBC, Vancouver, Canada

e-mail: ameryw@yahoo.com

The Canadian English Language Proficiency Index Program General (CELPIP-G) Test is a standardized assessment of English functional ability in working and community settings. The interpretation the CELPIP-G test scores are criterion-referenced to the 12-level Canadian Language Benchmarks (CLB) and used for Canadian immigration and citizenship purposes. Validity is vital to score interpretation and use when CELPIP-G is used for such high-stakes decisions. The purpose of this study is to examine the intended claims of the CELPIP-G in such that (1) the CELPIP-G scores reflect individuals' English functional ability and (2) higher functioning participants would be classified as having higher CLB levels as assessed by the CELPIP-G. The revised CELPIP-G Test was pilot tested on a sample of 350 voluntary participants who were living in Canada on various types of visa or were permanent residents and citizens. Participants were surveyed on their English language background (e.g., years of studying English) as well as how their current engagement in English at workplace and in the community (e.g., go shopping and reading work reports). Latent class analysis was conducted to identify groups of participants who differed in their ways of English language engagement using structural Equation Modeling (SEM). Test takers' English language backgrounds were then modeled as predictors for the latent classes, which, in turn, were modeled as predictors for CELPIP-G assessed CLB levels. Three latent classes were identified. The first class had little regular social and no work engagement. The second class engaged socially and in work settings other than office environments. The third class engaged both socially and in office environments. It was found that English language background predicted latent class membership and the latent class predicted CLB levels. The study concludes that SEM-based LCA is a strong method for developing warrants that support a validity argument.

Using Semi-Supervised Approach to Improve the Attribute Estimation in Cognitive Diagnostic Model

Huey-Min Wu[1], Bor-Chen Kuo[2], Chun-Hua Chen[2] and Wenchih Lin[2]

[1]National Academy for Educational Research, New Taipei City, Taiwan; [2]Graduate Institute of Educational Measurement and Statistic, National Taichung University of Education, Taichung, Taiwan;

e-mail: whm@mail.naer.edu.tw

According to the type of learning procedure used to generate the output value, there are two approaches, supervised learning and unsupervised learning in pattern recognition. In supervised learning, such as discriminate analysis, a set of training data, consisting of a set of instances that have been properly labeled by hand with the correct output, is provided .On the other hand, Unsupervised learning, such as cognitive diagnostic model, assumes training data that has not

been hand-labeled. Semi-supervised learning gets between unsupervised learning and supervised learning. Using a small amount of labeled data with a large amount of unlabeled data in semi-supervised learning . Taking DINA as an example, the semi-supervised learning is applied into the estimation algorithm in cognitive diagnostic model in this study. A simulation study is conducted to evaluate the performance of the proposed algorithms. A real data is taken as an example. The results showed that semi-supervise approach can increase precision in attributes correct classification rate under simulation study with different conditions. For the real data, there are totaling 286 subjects and 10 attributes. Using expert decision as criteria, the correct classification rates for unsupervised learning and semi-supervised learning are 68 % and 88% respectively under the real data. Based on the results, semi-supervise learning can increase performances of DINA model in estimating attributes.

Test Repurposing: Performance Statistics as Score Validity Evidence*

Nuo Xi, Maria Elena Oliveri and Christine Mills
Educational Testing Service, Princeton, NJ, United States
e-mail: nxi@ets.org

The objective of this study is to discuss fairness issues inherent in test repurposing. A test is repurposed if it is used for purposes or with populations differing from those for which the test was originally intended (Wendler & Powers, 2009). Foreseeable challenges arise when expanding the use of a test internationally that was developed for a single country. Such challenges include diversity in test takers' linguistic/cultural background and differences in curriculum exposure, which may introduce construct irrelevant variance and lead to threatening score comparability across countries. Therefore it is crucial to assure that the test measures the same construct across populations. The current study will illustrate an empirical approach to evaluate the test score comparability through a series of statistical analysis including differential item functioning, reliability estimates, correlations among sections, and analyzing anchor-total relationship for test takers from different countries. Additionally, we will interpret the results using the conceptual framework proposed by Oliveri, Lawless, and Young (2013), which focus on the validation of repurposed assessments. We use data from EXADEP, a test measuring verbal and quantitative skills of the applicants to graduate schools. Three sections of the test are in Spanish and one is in English. The test was originally developed to be administered in Puerto Rico, and now it is given in several Spanish speaking countries including Spain. Preliminary results comparing examinee performance across three countries (i.e., Puerto Rico, Colombia, Ecuador) indicate differential test performance on EXADEP, and the difference varies by sections (i.e., Verbal Aptitude, Quantitative, Written Expression, and English). Additional analysis as described above will be ready by the presentation date. The proposed study evaluates the fairness of a repurposed test through a series of statistical analysis, which serves as an empirical approach to enhance the development of fair repurposed assessments.

Dissecting Professional Guidelines for International Personnel Testing

Yongwei Yang[1], Kurt F. Geisinger[2], Anna Truscott-Smith[3] and Tzu-Yun Chin[2]
[1]Gallup, Omaha, NE, United States; [2]Buros Center for Testing, Lincoln, NE, United States; [3]Gallup, London, United Kingdom
e-mail: yongwei_yang@gallup.com

Personnel selection practices are increasingly global. Corporations are extending their presence to multiple countries and need selection tools and procedures that adapt to and service international settings. International exchanges also have encouraged companies to seek

personnel selection services generalizable beyond national borders. We see the importance of consistent professional expectations –common standards and guidelines –encouraging valid solutions and proper test use. Numerous professional standards are available. However the many details involved in adapting and validating instruments for different international contexts can be overwhelming. This paper summarizes the most relevant standards prescribed by major professional guidelines for international personnel testing. We synthesize their similarities and differences and provide clarifications and practical recommendations. We focus on two ITC guidelines (ITC Guidelines on Adapting Tests; ITC Guidelines on Test Use), the ISO 10667, SIOP's Principles for the Validation and Use of Personnel Selection Procedures, and the 2014 AERA/APA/NCME Standards for Educational and Psychological Testing. The ITC guidelines and ISO 10667 were developed particularly for international applications. We included the SIOP Principles because this document (like ISO 10667) targets personnel testing and is also known internationally. Finally, the AERA/APA/NCME Standards is perhaps the most widely cited document by international testing professionals, even though its actual scope is limited to the United States and Canada. We start with a summary of the background, purpose, scope, target audience and intended applications of the documents. Next, we synthesize the commonalities and differences among these documents. Finally, we discuss what these guidelines offer to address a number of practical questions commonly faced by personnel testing users, such as re-testing, language accommodation, and using mobile devices for test administration. The paper provides a reference to best practices. It also facilitates the discussions and debates about the principles for sound development and use of personnel tests in cross-national settings.

Situational Judgement Test (SJT) for Measuring Non-Academic Attributes of Malaysian University Students

Haniza Yon[1], Rosna Awang Hashim[2], Tengku Faekah[2], Tengku Ariffin[2], Kok Mun Yee[3] and Nur Ayu Johar[3]

[1]Mimos Berhad, Kuala Lumpur, Malaysia; [2]Universiti Utara Malaysia, Sintok, Kedah, Malaysia; [3]Mimos Berhad, Technology Park Malaysia, Kuala Lumpur, Malaysia

e-mail: hanizayon@yahoo.com

SJTs are instruments that present test takers with hypothetical scenarios and require them to indicate how they would or should respond in the situations described. SJTs represent a generic approach that can be tailored to measure different core constructs such as personality, values, cognitive ability, and knowledge. SJTs have been used as predictors and as criterion measures, and their interpretation as knowledge measures is consistent with both uses. They offer a standardised method for objectively assessing a broad range of non-academic attributes, and exhibit high face validity provided the scenarios included in the test are based on construct-relevant situations. In this study, two alternate forms of an SJT were developed and administered to 1800 Malaysian final-year university students in order to measure knowledge-based constructs deemed important for success at entry-level positions in various careers, including critical thinking and problem solving, enterprise, integrity, leadership and teamwork skills. In addition, the students completed a personality test based on the Big Five Model. The data were analysed using the item response theory (IRT) approach, and test equating was performed in order to put the scores from the two alternate forms on the same scale. The results showed significant differences among students in terms of enterprise, integrity, and teamwork, and in the personality traits of agreeableness and extraversion. Scores on the SJT test were also found to be significantly correlated with students' parental income and English language proficiency. The results provided evidence that the assessment met established criteria for many aspects of validity, though some other aspects remain to be investigated in the future. This case

study illustrates the important methodological challenges inherent in SJT development in the Malaysian context, challenges which must be overcome in order to maintain test standards.

The Effect of Removing Examinees with Low Motivation from Large Scale Assessment Data Calibration

Carlos Zerpa
Lakehead University, Thunder Bay, Canada
e-mail: czerpa@lakeheadu.ca

Every year, about 140,000 Grade-9 students in the province of Ontario, Canada, are given a large-scale assessment in mathematics to monitor student academic achievement and performance. Current item-response models (IRM) used to calibrate the test item parameters do not account for the effect of student low motivation. The purpose of this study was to examine the effect of removing examinees with low motivation on test item parameter estimates by using differential item functioning (DIF) techniques on large scale assessment data. A sample of 63,783 Grade-9 students was used for this study. Student motivation was identified from self-report data using a principal component analysis. Two components scores, math-value and interest, were computed for each examinee and merged with student item responses to create two groups, high-motivated and low-motivated examinees. These groups were used to examine the effect of low motivation on the estimates of test item parameters between a standard 3-parameter logistic (3PL) and modified 3PL IRM, which did not include examinees with low motivation. The data were analyzed using DIF techniques. The results from the DIF analysis suggested that some item parameters were overestimated when examinees with low motivation were included in the IRM calibration. The outcome of this study supports the literature and seems to provide an avenue to reduce item bias so that more valid interpretations of test results can be made from large scale assessment data. The outcome of this study may also have implications for educational agencies, teachers and policy makers to better assess student mathematics academic performance and make better curriculum decision changes in our current educational system.

Scoring Situational Judgment Tests with the Nominal Response Model

Jiyun Zu and Patrick Kyllonen
Educational Testing Service, Princeton, NJ, United States
e-mail: jzu@ets.org

A situational judgment test (SJT) item presents a situation and several possible responses to that situation. Most often examinees are asked to select the best (or also the worst) from the possible responses. A characteristic of SJTs is that they are often designed to measure judgments in situations for which there is not a consensus on the correct response, or there might be multiple correct or acceptable responses. In this paper, we evaluated the use of the nominal response model (Bock, 1972) as a scoring method for SJTs. The NRM is an item response theory model that does not require prior knowledge of an item key or ordering of the response options, but assigns data-driven partial-credit to each option. Through a secondary analysis of three large-scale datasets, we compared reliability and correlations with external variables of NRM scores with those from various conventional (i.e., number-correct and consensus scoring) and other item-response-theory scoring approaches. We found that NRM provides more reliable and valid scores than all other methods when the item key is ambiguous, although all scoring methods provide interchangeable scores when the key is clear. Category response curves based on NRM parameter estimates could reveal whether an item key is clear or ambiguous. We suggest

keeping options from critical incidents rather than imposing obvious keys in SJT development, and using NRM for scoring if item keys are suspected to be ambiguous. Datasets were a middle school sample (N>2000) with an SJT measuring emotional management, a college level sample (N>1000) with an SJT measuring teamwork and collaboration, and another college level sample (N > 5000) with an SJT measuring college performance.

Posters

An Investigation of The Prepredisposing Factors of Cheating in University Assessments

Bamidele Abiodun Faleye
Obafemi Awolowo University, Ile-Ife, Nigeria
e-mail: bamidelefaleye@yahoo.com

Cheating is any act that is capable of giving a candidate or candidates undue advantage over others in any form of assessment. This study investigated the magnitude and direction of predisposing factors of cheating in Obafemi Awolowo University, Ile-Ife, Nigeria. It also identified the causal factors of cheating by selected undergraduate students. The paper determined the relationship between the independent variables (causal factors) and the path of their relationship with the dependent variable (cheating behaviour). The paper investigated the contribution of each of the independent variables to propensity to cheat in the selected university assessments. The study adopted the ex-post-facto research design. The population for the study comprised all undergraduate students of Obafemi Awolowo University, Ile-Ife, Nigeria. A sample of 650 undergraduate students was selected using volunteer sampling. Data were collected using a battery of adopted scales. Data were analysed using percentage, analysis of variance, multiple regression and path analysis. Results showed that a number of factors predisposed students to cheating in the university. Students also differ in their involvement in cheating, with more males engaging in it than female students. It was also revealed that religiousity, parental socio-economic status, peer influence and adequacy of preparation by both lecturers and students contributed strongly to propensity to cheat. It is concluded that adequate teaching and positive study habits are essential factors capable of reducing students' predisposition to cheating in the university.

Converging Relational Orientation as a Constituent Dimension of Identity in Self-Descriptions and Sources of Identification in Self-Report: A Multimethod Approach

Byron Adams[1], Fons van de Vijver[2], J. Nel[3], Sumaya Laher[4], Johann Louw[5], Luzelle Naude[6], Florence Tadi[6] and Joey Buitendach[7]

[1]University of Johannebsurg, Auckland Park, South Africa;[2]The Netherlands North-West University, Netherlands; [3]North-West University, South Africa; [4]Witwatersrand University, South Africa; [5]University of Cape Town, South Africa; [6]Free State University, South Africa; [7]University of Kwazulu Natal, South Africa

e-mail: bgadams@uj.ac.za

Individuals negotiate identity intra- and interpersonally. We consider multimethod approach to our studyof relational orientation. Relational Orientation refers to the relational part of our identity andis defined as the perceived importance individuals or groups attach to personal, relational, and/or social aspects of their identity. It extends the individualism-collectivism (IC), and self-construal (independence-interdependence) dichotomies, by considering implicit relational orientation (humanitarianism/altruism) and explicit relational orientations (relations with significant others). Working in different ethnocultural groups in South Africa, 1134 participants (75.08 % females, Mage = 20.03 years, SD = 2.37): Black (n = 360), Coloured (n = 109), Indian (n = 62), and White (n = 603), completed an open format qualitative measure adapted from the Twenty Statements Test anda closed format Likert-type quantitative self-report measure with of statements relating to the respective categories. Using between-method triangulation, which is the use of contrasting methods to investigate convergence, we investigated whether our expectations with regards to relational orientation as originally developed in free self-descriptions would be confirmed in the sources of identification self-report measure. Results indicate that there is some convergence with regards to personal aspects of identity in both measures, whereas there was divergence with regards to broader relational

aspects. We discuss the scale development for multicultural societies, together with and implicationsfrom the South African context.

Using the Guo & Drasgow Z-Test to Detect Cheating in a Real Selection Context.

David Aguado[1], Francisco José Abad[2], Julio Olea[2], Vicente Ponsoda[2], Alejandro Vidal[2] and Beatriz Lucia[1]
[1]Instituto de Ingeniería del conocimiento, Madrid, Spain; [2]Universidad Autónoma de Madrid, Madrid, Spain
e-mail: david.aguado@iic.uam.es

The verification of the scores obtained by applicants under uncontrolled internet testing (UIT) administration is an area of great academic and applied interest, due to the presence of what is commonly called cheating. In order to detect this dishonest behavior, different methods have been proposed to determine whether candidate scores obtained in a controlled environment differ significantly from those obtained in an UIT administration. The study explores the efficiency of the Z-test proposed by Guo & Drasgow (2010) in a real recruitment and selection context. A sample of applicants (N = 424) completed an English proficiency computerized adaptive test (eCat Grammar; Olea, Abad, Ponsoda, & Ximénez, 2004) under the UIT mode. Subsequently, under a proctored administration, they completed a verification test. The verification test is a 10-item length eCat Grammar with some additional restrictions related to the initial theta and the items available for each candidate. Under the UIT administration theta estimates are significantly higher than those obtained in the controlled administration. In fact, Guo and Drasgow Z-test flagged as possible cheaters about a 10% of the candidates. Some further comments will be offered on the agreement of the Z-test results with other verification procedures, and on the behaviors (presence of test surrogates or use of unauthorized materials) behind the observed cheating.

A SAS/IML Macro for DINA (Deterministic Input, Noisy “and” Gate) Model

Cigdem Alagoz-Ekici and Allan Cohen
University of Georgia, Athens, GA, United States
e-mail: cigdemalagoz@gmail.com

Diagnostic classification models have the potential to provide useful information to students and teachers about specific knowledge structures, skills and problem areas. The DINA model is one of the simplest and most studied of these models. This study presents a SAS/IML macro for estimating the DINA model using an EM algorithm described by de la Torre (2009). The macro was implemented in SAS in part because of the facilities in SAS for data manipulation. The simulation study reported here was designed to investigate the accuracy of parameter estimates provided by the SAS program. Data were generated for a sample size of 2,000 examinees and a test length of 30 items. Items were assumed to measure one or more of five attributes. Slip and guessing parameters were set to 0.2. A total of 100 data sets were simulated. The joint distribution of the skills patterns were generated with equal probability from a multinomial distribution. The results for the simulated data, and the mean estimate, root of the mean squared error, and the empirical standard error across the 100 replications will be presented. In this paper, use of a SAS program is illustrated for estimating parameters of the DINA model. Using this code, it is possible to manipulate the matrices at each step of the algorithm. In this way, it is possible to examine the performance of the DINA model under different practical testing conditions.

Application of the Experiment of Psi Abilities in Sports

Laith Mahmood Muhammad Al Azawe
Iraq Olympic Academy, Baghdad, Iraq
e-mail: lmlaithwhite952@gmail.com

The "Application of the Experiment of psi Abilities in Sports" Is a study completed through field work on different sports activities from 2005 till 2010. Its main objective is to examine the current status and thus the near future of the players based on the problems that affect them and the results of the games consequently, and which the instructors do not find clear from the incorporeal side. The using of my precognition for the psychoanalytic diagnosis performed on sports games (8 handball games, 2 basketball games, 1 volleyball game and 1 fencing game for girls), focused on the level and quality status of the concentration, energy, tension, anxiety and the willingness to compete of the athletes teams for age 20-30. It was performed through visions of future images in my insight, coming forth in the form of microscopic waves, and symbolized in diagrams representing every game. The diagnostic results of all the games were successful, I found that these common variables among athletes coming from different backgrounds of sports activities are essential for winning the game and they affect the results whether it is positively or negatively. Finally, this experiment serves as a study on the possibility of use of psi abilities in the diagnosis of psychological and biological variables in human beings at different times during the game, as well as showing the role of psi abilities in sports games.

The Saudi Standardized Achievement Admission Test (SAAT): A study of Construct – Related Validity Evidence

Amjed A. Al-Owidha
King Fahd University of Petroleum & Minerals/ National Center for Assessment in Higher Education, Riyadh, Saudi Arabia
e-mail: A.OWIDHA@Qiyas.org

With the increased demand on the standardized achievement test scores as being a major component required for university and college admission in Saudi Arabia, it is obligatory for the National Center for Assessment in Saudi Arabia (NCA), as the SAAT developer, to continually check and maintain the quality of the SAAT test by collecting evidence to test validity interpretation and use of its scores. This study aims to collect construct – related validity evidence of the SAAT scores to support its interpretation and use. Over a thousand (approximately 10% of the participants who took the 9040 form of SAAT test in 2012) examinees were randomly selected and their responses used in this analysis. Two approaches were used, mainly, item response theory (the Rasch model) and confirmatory factor analysis of item parceling. First, Rasch analysis was applied to the SAAT test data. Several Rasch indices were used to evaluate the internal structure of the SAAT. Mainly, overall fit statistics, item fit statistics, person reliability index, point – measure correlations, principle components analysis of Rasch residuals, and uniform differential item functioning were all computed and used. Second, the confirmatory factor analysis of item parcels was applied to the SAAT test data. Several fit statistical indices were used to evaluate the fit of the measurement model to the SAAT data. More specifically, the Chi- square test (X2), the root mean square error of approximation (RMSEA), the Goodness of Fit Index (GFI), the comparative fit index (CFI), Tucker – Lewis index (TLI), the Standardized mean square residuals (SRMR) , and the Akaike Information Criterion (AIC) were all used to further investigate the internal structure of the SAAT. The findings of both paradigms lend support to the internal construct validity of the SAAT in terms of its score interpretations and use.

Development of a New Scale for Assessing Social Desirability

Gema Alonso, Ángel García-Pérez, Ignacio Pedrosa and Eduardo García-Cueto
Universidad de Oviedo, Oviedo, Spain
e-mail: aonsodiegogema@gmail.com

The use of psychological tests for the decision making on the different fields of people´s life is increasing, and thus, the requests on quality demanded are also greater. Therefore, it becomes crucial knowing and identifying the bias that could affect results interpretations obtained in the tests. For this reason, it is important to also recognize those subjects tending to fake the tests answers with the intention of giving a positive image of themselves. The aim of this current work was to develop a new measurement scale which would allow assessing social desirability in a brief way. On the other hand, the possible effect that sociodemographic variables could have over social desirability was studied. First of all, a pilot study to test the psychometric functioning of the items was made. Subsequently, the already mentioned scale was applied onto a new sample along with a sociodemographic scale in order to evaluate the socioeconomic status, find the study level, sex and the individual age. Then, the differences according to the defined sociodemographic variables were studied (CL=95%). With effect from pilot study, 9 items were deleted according to the index of difficulty and discrimination, leaving the scale made up with a total of 12 items. The results show that the developed scale contains appropriate psychometric properties. Finally there are no statistically significant differences on social desirability according to the assessed sociodemographic variables. Henceforth, it seems reasonable to think that these variables are not relevant when faking psychological tests' answers. Finally, a brief and accurate test is provided for the assessment of the social desirability on general population.

Impact of Pakistani Norms on Scores of Urdu Adaptation of WISC-IV

Saima Ambreen and Anila Kamal
National Institute of Psychology (NIP), Quaid-i-Azam University, Islamabad, Pakistan
e-mail: saima.ambreen.awan@gmail.com

The current study is aimed at exploring the impact of Pakistani norms on WISC-IV Scores. For that purpose scaled scores and composite scores of 220 children on Urdu adaptation of WISC-IV were compared when UK (Wechsler, 2004) and newly developed Pakistani norms are used. Results indicated a significant increase in composite scores of verbal comprehension, perceptual reasoning, and processing speed indices when Pakistani norms are used. Similarly, FSIQ increased substantially on using the regional norms instead of UK norms. Considering scaled scores, vocabulary has shown the maximum gains, while conversely digit span has shown little drop in scale scores when Pakistani norms are used. These gains have been further explored on three age groups of 6-8 years, 9-12 years, and 13-16 years. Probable reasons for these gains or drops have also been discussed. The findings strongly support the notion of establishing regional/national norms for assessing constructs like intelligence especially for children.

Validation of the Condom Use Self-Efficacy Scale to the Brazilian Context: Factor Structure, Reliability, and Validity

Josemberg Andrade[1], Felipe Valentini[2], Nelson Hauck Filho[3] and Kaline Silva Lima[1]
[1]Federal University of Paraíba, João Pessoa, Brazil; [2]Universo University, Niteroi, Brazil; [3]São Francisco University, Itatiba, Brazil
e-mail: josemberg.andrade@gmail.com

Risk behaviors represent a major target in health public strategies that seek to prevent AIDS. In order to deepen the theme, the present research aims to adapt the Self Efficacy Scale in Condom Use (CUSES) to the Brazilian context. The CUSES consists of four dimensions: Appropriation, Assertiveness, Pleasure and Drugs, and STDs. We compared the means among Brazilians from five geographic regions. We assessed, by means of an online survey, 1,215 subjects from the Northeast (38.9%), Southeast (25.8%), Midwest (16.5%), North (7.2%), and South (11.2%). Most participants were female (65.4%), varying in age from 15 to 62 years (M = 23.9; SD = 6.5), 69.8% were single, and 27.6% were engaged in a couple relationship; 32.9% reported always usingcondom, whereas 27.8% often use condom, 20.7% sometimes, and 13.8% seldom. We performed a confirmatory factor analysis (CFA) of data, followed by analyses of variance (ANOVAs). Considering the factor model proposed by Asante and Doku (2010), we retained four factors for 14 items. The results supported the factorability (KMO = .77; X^2 [91]= 6368.3; p< .0010), and the four factors explained 63.5% of variance–eigenvalues from 1.25 to 4.14. CFA goodness-of-fit was: X^2/df = 8.72; GFI = .91; AGFI = .89; CFI = .91; RMSEA= .08 Pclose = .000 (CI 90% = .74-.86). The reliability was assessed by Cronbach's alpha for the dimensions Appropriation (.75), Assertiveness (.79), Pleasure and Drugs (.54), and STDs (.86). Regarding the Anova, we did not find any significant difference betweengeographic regions. In conclusion, results suggest the CUSES is a valid and reliable instrument. Future studies should further elaborate interpretation norms for scores of the instrument, as well as estimate item parameters using Item Response Theory models.

An Investigation of the Six-factor Personality Structure Across Five Cultural Groups: Through the Lens of ESEM**

Ion Andrei[1], Dragos Iliescu, Kattiya Ratanadilok, Neeti Rana, Ari Widyanti and Said Aldhafri

University of Bucharest, Department of Psychology, Bucharest, Romania

e-mail: andrei.ion@fpse.unibuc.ro

The debates surrounding the universality of different personality models have encouraged scientists to seek new theoretical frameworks, as well as new data analysis techniques in an attempt to generate more coherent and replicable results. In terms of personality frameworks, one such endeavor was undertaken by Ashton & Lee (2001, 2004) resulting in a six-factor framework of personality. The most popular instrument used for measuring the six factors is the HEXACO Personality Inventory (Lee & Ashton, 2004). Some of the most widely used data analysis frameworks in personality research, Exploratory Factorial Analysis (EFA) and Confirmatory Factorial Analysis (CFA) have been criticized (e.g. Aluja, Blanch, & Garcia, 2005; Block, 2010). Exploratory Structural Equations Modeling (ESEM) is a relatively new, alternative method of data analysis that builds on the main features of EFA and CFA, but providing a more flexible analytical framework.

The current research aims at exploring the personality structure as measured with HEXACO, across samples from Hindi, Indonesian, Omani, Romanian, and Thailand cultural groups, by employing: CFA, EFA and ESEM.

The data used in this paper was obtained as part of the cultural adaptation process for the HEXACO across Hindi, Indonesia, Omani, Romanian and Thailand cultures. The sample sizes of the groups varied between 210 and 482 participants.

The EFA has generated different factorial structures across each investigated group. The structures obtained via ESEM exhibit a superior goodness of fit across the five investigated groups compared to the fit indices obtained via CFA. Even for those groups where the

goodness of fit obtained via ESEM is poor, the ESEM yielded a significantly better fit to the data, compared to the CFA-derived indices.

New Types of Test Items in Mathematics for Dichotomous IRT Models

Sayaka Arai

The national center for university entrance examinations, Tokyo, Japan

e-mail: sayarai@rd.dnc.ac.jp

In the area of educational measurement, item response theory (IRT) is used in many testing applications. While IRT provides several advantages over classical test theory, it requires strong assumptions. One of the assumptions is local independence: the items in a test should not be related to each other. However, there are many test items which are not locally independent. For example, items within a testlet are not necessarily independent each other, since they share common topic. Such sets of items are typically seen in reading comprehension test and mathematics test. In this study, testlet items in mathematics are dealt with. Usually, many of items within a testlet in mathematics are not independent, since previous items need to be answered correctly to solve later items. Therefore it is said that testlet items in mathematics should not be analysed with dichotomous IRT models but with polytomous IRT models or other models. However, new types of test items, which are called sequentially-presented items, were proposed by Okubo in 2013. They are also a set of items in mathematics while items are not dependent each other. The purpose of this study is evaluating sequentially-presented items. New items were developed and an experiment was conducted to measure their availability. In this experiment, several students were video-taped while they were working on these test items. The videotape was analysed and the results suggested that test items within sequentially-presented items were surely not dependent each other but the difficulty of some items were low compared to testlet items.

Factor Validity of Torrance Test of Creative Thinking-Verbal form B in Argentinean Students

María Aranguren[1], Gabriela L. Krumm[2], Vanessa Arán Filippetti[2] and Viviana Lemos[2]

[1]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Centro de Investigaciones en Psicología y Psicopedagogía (CIPP). Pontificia Universidad Católica Argentina (UCA), Ciudad Autónoma de Buenos Aires, Argentina; [2]Centro Interdisciplinario de Investigación en Psicología Matemática y Experimental Dr. Horacio J. A. Rimoldi (CIIPME). Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), República Argentina. Universidad Adventista de Plata, Ciudad Autónoma de Buenos Aires, Argentina

e-mail: maria.aranguren@yahoo.com

The Torrance Test of Creative Thinking (TTCT) battery includes a Figural and a Verbal version. Although the TTCT is one of the most widely used test of creativity (Colangelo & Davis, 1997), its' construct validity remain being questioned and analyzed (Kim, 2011). There are few studies of the TTCT- Verbal factor structure (e.g. Krumm & Lemos, 2010). In contrast, several recent studies of the TTCT factor structure have been done using the Figural version. The TTCT-Verbal consists of six activities which evaluate three creative skills: (a) fluency, (b) originality and, (c) flexibility (Torrance, 1990). These abilities are scored in every activity resulting in three subscales that were proposed by Torrance (1990) in the TTCT scoring system. The purpose of this research is to analyze the factor structure of the TTCT-Verbal Form B in Argentinean students. Based on previous studies, two different models were compared through confirmatory factor analysis (CFA). The first model proposes six factors which represent the six activities with their respective indicators –fluency, flexibility and originality-. The second model consists of

three factors that represent the three skills assessed, namely: fluency, originality, and flexibility, and the corresponding indicators for each variable (e.g. to the fluency variable corresponds the fluency indicator of each activity). Respondents consisted of 432 Spanish speaking youngsters of both genders aged 15 to 26. CFA were conducted to test the fit of the two models. According to the research findings, the model which showed the most satisfactory fit identifies six correlated factors (Chi-square = 414.48; df = 116; Chi-square/df = 3.57; GFI = .90, NFI = .95; CFI = .96 and RMSEA = .077). Considering the six factors found in this research, it would be advisable to reconsider the use of the ability scores as separate factors.

BAT-7, TEA Abilities Battery: IRT/Ordinal Reliability and Internal Structure

David Arribas-Aguila
TEA Ediciones, Madrid, Spain
e-mail: david.arribas@teaediciones.com

BAT-7 is a new psychometric tool for assessing three intellectual domains, general factor (g), fluid intelligence (Gf) and crystallized intelligence (Gc), and eight cognitive abilities: Verbal (V), Spatial (E), Attention (A) and Concentration (CON), Reasoning (R), Numerical (N), Mechanical (M) and Orthography (O). It consists of three forms of increasing difficulty (E, M and S) and it is focused on the assessment of school and college students, as well as adults with different educational background. The present study aims to collect the data necessary to estimate IRT parameters, to determine ordinal scores reliability and to gather validity evidences based on its internal structure. BAT-7 was administered to a Spanish standardization sample of 4,263 students and 1,507 adults. A Bayesian adaptation of the MML estimation technique was used in a 3 parameter logistic model and model-data fit was examined using standardized residuals and chi-square statistics. Reliability was studied by means of ordinal alpha and tests information functions. Validity was explored through multigroup CFA. IRT model showed statistical fit for all items. Ordinal alpha ranged from .79 to .91 for test scores and from .91 to .97 for composite scores. The model based on CHC theory showed a good fit to the observed data (RMSEA=.034; SRMR=.026; CFI=.981). Psychometric properties analyzed in terms of IRT modeling, reliability and validity evidences support the usage of the battery for applied purposes. BAT-7 can be described as a high informative and reliable measure of the most important cognitive abilities. Likewise, SEM analysis seems to support a hierarchical model reflecting the g, Gf and Gc factors of the CHC theory.

Factor Structure and Reliability of the Spanish Workgroup Emotional Intelligence Profile-Short version (SWEIP-S) in a Sample of Primary School Teachers

José María Augusto-Landa, Manuel Pulido-Martos and Maria Luisa De La Casa-Pegalagar
University of Jaén, Jaén, Spain
e-mail: jlanda993@gmail.com

Group Emotional Intelligence (GEI) refers to the development of a collective EI resulting from interactions of a workgroup (Ghuman, 2011). EI as a group construct should be evaluated with instruments that use group as unit of analysis (Klein, Dansereau, & Hall (1994). The Workgroup Emotional Intelligence Profile (WEIP, Jordan, 2000; Jordan, Ashkanasy, & Hooperb Härtelb, 2002) is an instrument of this type. A short version of WEIP has been adapted for the Spanish population (López-Zafra, Pulido-Martos, Augusto-Landa & Berrios, 2012). EI research in populations of teachers is extensive (Brackett & Caruso, 2007; Fernández-Berrocal & Ruiz-Aranda, 2008). In order to analyze EI from a group level, it's necessary to first determine the validity and reliability of instruments like SWEIP-S, in this population. To assess the original

factorial structure of SWEIP-S in sample of teachers, and to analyze the reliability of the instrument. A cross-sectional design was used. Data were collected from a sample of 453 teachers (275 women) from different schools in Jaén (Spain). A confirmatory factor analysis (CFA) was performed to examine the scale's construct validity. Reliability coefficients (Cronbach's alpha) were .86, .77, .82 and .88 for awareness of one's own emotions, management of one's own emotions, awareness of other's emotions and management of other's emotions, respectively. A series of CFA were performed to find a better-fitted model. The final model yielded a four-factor structure with two error correlations (X^2 S-B = 114.14; CFI = .98; NNFI = .98; RMSEA = .028 [.000-.045]). All factor loadings were significant ranging from .64 to .84. The four factor structure with two error correlations of SWEIP-S is valid and reliable in the sample of primary school teachers.

Overcoming Unbalanced Item-Exposure Rates in Computer Adaptive Testing with Weighted Item Information Functions

Alexander Avian and Andrea Berghold

Institute for Medical Informatics, Statistics and Documentation; Medical University of Graz, Graz, Austria

e-mail: alexander.avian@medunigraz.at

In computer adaptive testing (CAT) unbalanced item exposure rates is a well-known problem. A similar phenomenon can be encountered in clinical trials where e.g. complete randomization can lead to unbalanced groups. One successful strategy for balanced randomization is biased coin randomization. We investigated the adaptability of the concept of biased coin randomization for CAT and the performance on item information functions in simulation studies. Different scenarios based on the "amount of bias" were compared to a model without exposure rate control with respect to reliability, root mean square error (RMSE), bias and mean absolute error (MAE) to evaluate the performance of a "weighted item information function" (wIIF). In the wIIF model the item information function is weighted by an inverse factor (fn). This factor increases by a predefined constant (g) times the number (s) an item has been selected (fn=f0+g*s). This was done with different selection rules for the starting item and test termination rules. In the wIIF model a reduction in the maximum item exposure rate, median item exposure rate and number of items never exposed was found depending on the "amount of bias" which was introduced. In spite of a more balanced exposure rate there was no difference in bias, RMSE and reliability. Different selection rules of the starting item and test termination rules resulted in comparable results. For the MAE lower constants ($g<.5$) resulted in a bigger MAE. For tests with a low number of items ($n<15$) the MAE was bigger for higher g ($g>.5$). If the constant g is higher than .5 simulation studies for wIIF leads to a more balanced item exposure rate without reducing precision. For low number of items the wIIF results in comparable RSME, bias and reliability but higher MAE.

Spelling Test as a Tool for a Diagnosis of Risk of Dyslexia

Elzbieta Awramiuk[1], Grazyna Krasowicz-Kupis[2], Katarzyna Maria Bogdanowicz[2], Dorota Kwiatkowska[2] and Katarzyna Wiejak[2]

[1]University of Bialystok, Bialystok, Poland; [2]Educational Research Institute, Warsaw, Poland

e-mail: eawramiuk@poczta.onet.pl

The presented study is a part of IBE Dyslexia Project, which concerns the early identification of specific reading and spelling disorders, specifically the risk of developmental dyslexia. The main goal is to develop a battery of tests to measure reading and spelling abilities in children before starting and at the beginning of formal reading instruction, that are predictive of these disorders.

The aim of this study is to examine an assessment tool to test the graphotactic awareness as an important component of invented spelling. The purpose is to diagnose whether the child has a sense of legal letter combinations in Polish. The task involves assessing pseudowords. In each pair is a string of characters, that are possible in Polish (eg, TAB, DYK, AMA) and the incredible string of letters (BTA, YDK, AOA). With each pair, child has to decide which string of letters is likely. The test consists of 30 items and was tested in a sample of 400 children, at the beginning of formal reading instruction (aged 6-8). The psychometric criteria show adequate values for reliability. Further results from a pilot study and additional aspects of validity will be presented in this study.

Single Factor of Personality: Neither Universal, nor Useful

Rob Bailey[1], Leonie Nicks[2] and Fiona Young[2]

[1]OPP Ltd, Oxford, United Kingdom; [2]Oxford University, Oxford, United Kingdom

e-mail: rob.bailey@opp.eu.com

The Single Factor of Personality (SFP) is the idea that personality can be summarised most parsimoniously not by 5 big factors, but by just one. Supporters of the SFP claim it predicts overall work performance. This work investigated SFP claims, particularly whether the SFP, a 5 Factor model, or a 16 Factor model would be a better predictor of work engagement, work performance, salary and promotion. Several datasets were analysed: 1. UK and Republic of Ireland (N 1,212), 16PF5 and criteria questions, including work outcomes; 2. Personnel samples: (UK N 15,000 males and N 15,000 females; US N 30,567) where 16PF had been used for recruitment and development; 3. US manager dataset (N 279): individuals on a management course completed the 16PF5 and the Benchmarks 360 assessment of competencies. Analysis included PCA Exploratory Factor Analysis to look for the SFP and linear multiple regression analyses to look for relationships between personality models and Impression Management with work performance and engagement. Multiple attempts to create a Single Factor of Personality failed. EFA consistently produced a two factor solution. Regression analysis showed that the two factor solution was poor at accounting for work-based criteria, the 5 factor solution fared slightly better, and the 16 factor model was the best model. For example, a two factor model ($R=.151$), a 5 factor model ($R=.252$) and a 16 factor model ($R=.312$) all significantly predicted salary ($p <.001$). However, only the 16 factor model could predict all of these criteria: salary, boss rating of performance, and engagement at work. SFP is not found in 16PF, so cannot be considered a universal personality structure. Use of broad factors in research is likely to underestimate the role of personality; their use in high-stakes applications such as recruitment would be of great concern.

Factor Structure of the Scale of Sense of Community in University Online Courses

Giulia Balboni[1], Stefano Cacciamani[2] and Vittore Perrucci[2]

[1]University of Pisa, Department of Surgery, Medical, Molecular & Critical Area Pathology, Pisa, Italy; [2]University of Valle d'Aosta, Aosta, Italy

e-mail: giulia.balboni@med.unipi.it

In online courses, Sense of Community (SC) affects the quality of students' learning and satisfaction. Nevertheless, there are no tools that allow us to measure in online courses the four dimensions of the main reference SC model (McMillan & Chavis, 1986) - membership, influence, sense of personal fulfillment and integration of needs, and shared emotional connection. For this purpose, the Scale of Sense of Community in online Courses (SSCC) (Perrucci et al., 2012) has been developed according to the SC model. The aim was to verify the factor structure of the

SSCC. 552 students (86% females) of university online courses provided by several Italian universities completed the SSCC. The questionnaire was made up of 60 items rated on a 4-Likert scale, which allowed for the measurement of each of McMillan and Chavis' model dimensions. An exploratory factor analysis (principal axis factoring) was carried out with an oblique rotation (statistical assumptions were confirmed). Three factors were extracted based mostly on the interpretability of the pattern of the factor loadings and the scree plot. For each factor, items with a factor loading higher than .30 were selected; in this way, 43 of all the 60 items were retained. The three factors were labeled as Membership (20 items), Fulfillment of needs and goals achievement (14 items), and Mutual influence between individual and community (9 items). Total variance explained was equal to 35.28%. The factors' inter-correlation ranged between .43 and .49. A confirmatory factor analysis will be undertaken in a different sample of 233 university students (74% females) to evaluate the accuracy of the three factors' structure. Parcels of items will be used. The factorial structure identified may allow for the planning of targeted interventions on each of the three SC components.

Diagnostic Efficiency Statistics of the Diagnostic Adaptive Behavior Scale

Giulia Balboni[1], Marc J. Tassé[2], Robert L. Schalock[3], Sharon A. Borthwick-Duffy[4], Scott Spreat[5], David M. Thissen[6], Keith F. Widaman[7], Dalun Zhang[8] and Patricia Navas[2]

[1]University of Pisa, Pisa, Italy; [2]The Ohio State University, Columbus, OH, United States; [3]Bob Schalock & Associates, Chewelah, WA, United States; [4]University of California-Riverside, Riverside, CA, United States; [5]Woodland Center for Challenging Behaviors, Allentown, NJ, United States; [6]University of North Carolina - Chapel Hill, Chapel Hill, NC, United States; [7]Unviersity of California - Davis, Davis, CA, United States; [8]Texas A&M University, College Station, TX, United States

e-mail: giulia.balboni@med.unipi.it

Based on the diagnostic manuals, to be eligible for a diagnosis of Intellectual Disability (ID), an individual must function at approximately 2 SDs below the mean on one of the three Adaptive Behavior (AB) domains of conceptual, social, or practical or on the Total AB score. Therefore, the Diagnostic Adaptive Behavior Scale (DABS) must be valid and reliable to assess AB skills around the 2 SD cutoff. The aim of the study was to determine the accuracy of the DABS to correctly identify, across its three age forms, persons with diagnosis of ID and differentiate them from individuals without ID. Participants were the normative sample of the DABS made up of 1,058 persons, 4 to 21 years old (M = 11.1, SD = 4.9), 51% male. Of them, 12% had a formal diagnosis of ID; the remaining had typical development or other verified conditions which were different from an ID-related diagnosis. DABS accuracy was investigated: (1) comparing the SS of individuals with and without ID on each domain and Total DABS score; estimating (2) discriminant ability (sensitivity, specificity, positive and negative likelihood ratios [LR+, LR-]) and (3) predictive ability (Area Under the ROC Curve [AUC]) in correctly classifying individuals with and without ID based on their DABS SS. For the three DABS Forms: (1) Total AB SS and SS from three AB domains of individuals with ID were significantly below the SS of those without ID (effect sizes > 2.20); (2) efficiency statistics coefficient ranges were .81-.98 for sensitivity, .89-.91 for specificity, 7.77-9.65 for LR+, and .02-.20 for LR-. AUC ranges of Total AB and of domains were .92-.98 and .87-.99, respectively. All the three DABS age Forms result in the accurate classification of 4-21 years old individuals with and without a formal diagnosis of ID.

Evidence of Validity for a Three-Factor Scale of Negative Post-Coital Emotions

Heitor Barcellos Ferreira Fernandes[1], Leif E. Ottesen Kennair[2], Claudio S. Hutz[1], Jean C. Natividade[1] and Daniel J. Kruger[3]

[1]Federal University of Rio Grande do Sul, Porto Alegre, Brazil; [2]Norwegian University of Science and Technology, Trondheim, Norway; [3]University of Michigan, Ann Arbor, United States

e-mail: heitor.barcellos@ufrgs.br

Little has been published about the prevalence, etiology, and phenomenology of Negative Post-coital Emotions (NPEs). However, recent scientific interest in NPEs and in their nomological net has greatly increased, but validated forms of measuring NPEs are lacking. We aimed to develop a comprehensive instrument encompassing all specific negative emotions shared by scientific classifications of emotions to test whether a convergent and theoretically-meaningful factorial pattern of NPEs in different cultures is found. Participants (N = 1399) from the Midwestern US, Norway, and Brazil (around 70% females; ages 18 to 30 to maintain hormonal-level uniformity; mean age 21.5) responded to two five-point Likert scales (where, for frequency, 1 represented 'never' and 5, 'all the time'; for intensity, 1 represented 'no intensity' and 5, 'extremely intense') containing seventeen NPEs and the instruction to recall the emotional experiences that immediately succeeded all past sexual activities with partners. Principal Axis Factor analysis with Direct Oblimin rotation produced three factors according to Horn's parallel analysis and the Kaiser-Guttman 'eigenvalues greater than one' rule, both for frequency and for intensity of NPEs, with the same emotions comprising each factor in both forms. High (>.90) coefficients of comparability were found for the three factors among the three nation-specific samples, and internal consistency for all factors ranged from .74 to .88. The three factors relate to (1) having a lesser and (2) a greater perceived desire for bonding and commitment than one's sexual partner does, and to (3) the maintenance of social reputation. The instrument was able to differentiate men from women, single individuals from those in a committed relationship. Most significant correlations between the three factors and validated scales measuring theoretically-related constructs from the Sexual Strategy Theory were as predicted by evolutionary psychology. Overall, the results provide adequate evidences of validity and precision for the three-factor NPEs scale.

Positive Strenghts to a Healthier Life - Optimsim and Life Satisfaction in a Brazilian Sample

Micheline Bastianello, Juliana Pacico and Claudio Hutz

Federal University of Rio Grande do Sul, Porto Alegre, Brazil

e-mail: mbastianello@hotmail.com

The aim of this study is to evaluate the relationship between optimism and life satisfaction in a Brazilian sample. Life satisfaction is the cognitive component of subjective well-being and refers to the assessment that individuals make about the quality of their lives. Previous research has found that life satisfaction is positively related to positive functioning and negatively related to distress or negative emotions. One of the most significant estimator that positively influence life satisfaction is optimistic thinking. Optimism can be defined as a stable personality trait related to positive expectations regarding future events. Several studies across cultures show a strong correlation between these constructs and their importance to a healthy life, in order to prevent disease and promote healthier behaviors. The sample of this study included 524 undergraduate students, aged 17 to 36 years (M = 21, SD = 3.2), 57% women. All participants completed Diener's life satisfaction scale and the Life Orientation Test (LOT-R), a scale that assesses dispositional optimism. Both scales were adapted and validated for use in Brazil. The sample was chosen for convenience and participation was voluntary. The correlation between optimism and

life satisfaction (r = .40) in the Brazilian sample is in agreement with the findings of the literature. Results indicate that, regardless of cultural differences between Brazil and North America (where those scales were originally constructed), the associations between optimism and life satisfaction are replicable, the results of studies are comparable, and the same instruments can be used for transcultural studies.

Translation and Validity Evidences to (EDS-21) for Brazilian Dancers

Andressa Melina Becker Da Silva[1], Claudiane Aparecida Guimarães[1], Isabella Goulart Bittencourt[2], Murilo Fernandes De Araújo[1], Renan De Morais Afonso[1], Tatiane Stephan Rocchetti Luz[1], Luiz Ricardo Vieira Gonzaga[1] and Sônia Regina Fiorim Enumo[1]

[1]PUC - Campinas, Uberlândia - MG, Brazil; [2]Universidade Federal de Santa Catarina, Uberlândia, Brazil

e-mail: claudianeguimares@yahoo.com.br

The exercise dependence composes a psychopathological condition in which the individual exercise excessively, accompanied by physiological and psychological symptoms. Has been evaluated in athletes by Exercise Dependence Scale (EDS-21). Can also occur in dancers that rehearse and compete for continuous hours, but there aren't instruments for this population. This study translated and adapted the EDS- 21 for Brazilian dancer's adolescents, finding validity evidences. The translation of six bilingual judges was taken, with little divergence, forming a synthesis of translations. We proceeded to back translation, with small differences that were detected by the original authors of the instrument, which were corrected. Cultural changes were necessary, given the specificities of the Brazilian dancers. In pilot testing, the instrument was applied in 35 dancers and is then applied to 338 dancers, most women (92.6 %), aged 10-19 years (M = 15.52, ± 3.55), who danced there 8.13 years (± 3.87). The application occurred during the international dance festivals held in Brazil. We checked for validity evidence of internal consistency, and the exploratory factor analysis using principal components with Promax rotation, in addition to calculating Cronbach's alpha. Factor analysis demonstrated the existence seven factors (Tolerance, Withdrawal, Intention Effect, Lack of Control, Time, Reductions in Other Activities, Continuance) - that explain 71.6 % of the total variance. The composition of the items on the factors demonstrated semantic and theoretical coherence, as expected, but the grouping of the factors was not equal to the original version of the instrument validation. Cronbach's alpha was 0.891, whereas question 12 ("I think about exercise when I should be concentrating on school-work") was deleted to decrease the accuracy of this release. It is concluded that this adaptation of EDS-21 can be used by Brazilian dancers and emphasizes the importance of attending to cultural issues in the validation of instruments.

Psychometric Analysis of the Eudemon Scale of Well-being in Brazilian Adolescents

Livia Maria Bedin Tomasi, Miriam Raquel Wachholz Strelhow, Luciana Rubensan Ourique Masiero, Bibiana Ramos Dos Santos, Marco Antônio Pereira Teixeira and Jorge Castellá Sarriera

UFRGS, Porto Alegre, Brazil

e-mail: liviabedin@gmail.com

The study of well-being is part of the positive psychology area which is dedicated to preventive and protective factors. Most studies in this topic are realized with adults. To extend the studies on well-being to adolescents, adapted and validated instruments for this age group are needed, considering their psychosocial development and cultural context. This study aims to evaluate the psychometric properties of the Eudemon Scale of Personal Well-being (EBP), with a sample of Brazilian adolescents. Participants are 1,479 adolescents with ages between 12 and 16 years old (M = 14.12, SD = 1.26), 64.7% girls and 35.3% boys. Reliability analysis, exploratory (EFA) and

confirmatory (CFA) factor analyses were carried out to verify the internal consistence and the scales factorial structure. The sample was divided to perform the EFA (n = 712) and the CFA (n = 767). Multigroup analyses were also performed to verify the configural and metric invariance regarding gender. The EFA indicated the presence of two factors, one related to the presence of well-being and other one to its absence. The instrument showed adequate levels of reliability for the two dimensions (α = .795, and α = .884 respectively). The CFA showed moderate fit indices (CFI > .90, RSMEA < .08) of the instrument with 21 items, after that three items of the original scale were removed. Multigroup analyses indicated that the factor structure of the scale and the factor loadings of the items are equivalent for boys and girls. Positive and significant correlations of the EBP were found with other instruments that assess well-being, through concurrent validation by Pearson correlations: Personal Well-being Index-School Children, Brief Multidimensional Student Life Satisfaction Scale, and two single-item scales (overall life satisfaction and happiness). It can be concluded that the instrument showed satisfactory psychometric properties, and can be used with Brazilian adolescents.

Innovative Approaches to e-Assessment: Possible Answers to the Guessing Problem in Multiple Choice Tests

Lenka Belanova and Hynek Cigler

Faculty of Social Studies, Masaryk University Brno, Brno, Czech Republic

e-mail: belanova@mail.muni.cz

Computerized testing has considerable potential not only to ease assessment load, but also to provide innovative and powerful modes of assessment. While the steady pressure for the use of more sophisticated question types in e-assessment is present, Multiple Choice (MC) is still the most frequently used question type. The conventional scoring method for MC tests is Number right, where the item score is based solely on the alternative selected by the examinee: full marks are given for the correct answer, zero points for the wrong or no answer. The goal of assessment is to accurately measure students' true ability. Nevertheless, under the Number right scoring scheme, the test evaluator cannot distinguish between correct answers based on knowledge versus those derived from a lucky guess. Since guessing adds random error to the variance of test scores and decreases both reliability and validity, it is a main factor to consider when attempting to improve MC tests. Several possible solutions to the guessing problem have been developed (in the CTT framework) and can be grouped into three main categories: Correction for guessing (also called Formula scoring), Confidence weighting (with Certainty-based marking scheme as its special case), Probability measurement. We present in detail the rationale of their scoring schemes as well as their advantages and disadvantages, and illustrate their psychometric properties on model data. While Probability measurement is the most appealing method from the psychometric point of view, the implementation of Certainty-based marking scheme into computerized multiple choice testing seems to be the most feasible choice. Its contribution to the increase of both reliability and validity has been shown, the formative potential has been advocated, while the marking scheme is still fairly simple and comprehensible to both students and teachers.

Depression from a Distance: The Use of Keystroke Dynamics in Tele-Diagnosis

Dennis Bernstein

PsychTech Ltd, Jerusalem, Israel

e-mail: dbern@psychtech.co.il

Over the past years digitization and telemedicine have seen vast improvements in resolution and accuracy of medical files and diagnostics. Over the same period advances have been made for evaluating user characteristics through background logging. Behavioral Biometrics has become a mainstream science, mostly in the field of security, such as identity verification and authentication. As more services are migrating to the cloud so has the field of medical diagnostics. Until now this has involved overt data collection such as questionnaires or observation formats. The current research is a pioneering attempt to collect data from a distance for the assessment of mental disorders through keystroke dynamics, a non-invasive and non-threatening method. 138 subjects were recruited through an internet survey company and requested to complete a series of textual tasks. The internet administered tasks were balanced as to comprise structured vs. unstructured stimuli, emotionally laden vs. emotionally neutral stimuli, copy vs. free text tasks and graphic vs. textual tasks. Keystroke logs were collected on variables including key-down time, flight time, backspace, delete, and other variables, for each task, producing a set of 69 features. After completion of the experimental tasks the subjects then completed the BDI (Beck Depression Inventory). Using ML (machine learning) a classifier algorithm was found to be accurate at the .8 level and a regression correlation for keystrokes and the BDI at .65. The use of non-intrusive data collection in medical and psychological diagnostics will play an important role in future telemedicine as accuracy and reliability of these technologies increase. As medical monitoring has moved from intrusive lab collected data, to ambulatory daily data collection so a tele-medicine behavioral biometric approach may prove useful for screening and monitoring of Psychiatric conditions.

Evaluating Sample Size Requirements for Calibrating Partial Credit Items

Joe Betts[1] and Doyoung Kim[2]

[1]Pearson Vue, Chicago, United States; [2]National Council of State Boards of Nursing (NCSBN), Chicago, United States

e-mail: joe.betts@pearson.com

As the field of testing moves to utilize the computer for assessment tasks, the range of item types at the test developer's disposal will increase. Many of the new, innovative item types will allow for a greater number of scoring options, i.e. multiple correct items, ordered response items, etc. With this move to utilize newer item types, some options will allow for multiple correct responses, for which partial credit might be given to any proper subset of the answer key. However, little research has investigated the adequate sample size requirements for evaluating these polytomous item response models. This research will simulate a number of important factors related to item calibration in order to provide some guidance on the minimal sample sizes needed to get stable parameter estimates from polytomous models. The factors that will be manipulated will be the distribution of ability with three levels corresponding to a normal distribution, a leptokurtic distribution (to simulate professional and licensure testing situation where the ability distribution is more compact), and a platykurtic distribution (to simulate an educational setting where students from a wide range of maturational levels might be responding to common items); the number of threshold per item ranging from one (to simulate the dichotomous model) to four (to simulate an item with five possible score categories); the number of responses per item ranging from 100, 200, 500, 800, and 1000; and the range of internal distances between the

category thresholds ranging from 1.0, 0.75, 0.5, 0.25, and 0.10 logits between the threshold separating score categories. Results will provide a practical basis for estimating sample sizes for obtaining stable estimates of item parameters for ordered response, partial credit models. Implications will be discussed and future research directions will be highlighted.

Career Planning Beliefs and Attachment Style as Predictors of Seeking Career Counseling

Hedva Braunstein-Bercovitz
The Academic College of Tel Aviv-Yaffo, Tel Aviv-Jaffa, Israel
e-mail: hedvab@mta.ac.il

There is strong evidence to support the effectiveness of career counseling (e.g., an enhanced sense of self-efficacy and career self-concept crystallization). However, a small portion of those who experience difficulties in career planning seek help and utilize career counseling services. The purpose of the current study was to examine how career planning beliefs and attachment style predict help seeking for career counseling. Career planning beliefs were conceptualized according to the principles of the Health Beliefs Model (HBM), which explains individuals' motivation for seeking help, and comprises five distinct components reflecting attitude towards seeking help. The objectives of the study were 1) to develop a new instrument that measures career planning beliefs, based on the five components of HBM, and 2) to examine the relationship between career planning beliefs, attachment style (secure, anxious and avoidant attachment), and help seeking for career planning. Two-hundred and two college students completed the following measures: (a) parental attachment style; (b) career planning beliefs (developed for the current study, and composed of five sub-scales, reflecting the five components of HBM, which were adapted for career planning by two licensed career counseling psychologists); and (c) readiness to seek career counseling help. A confirmatory factor analysis of the career planning beliefs inventory confirmed the existence of five major factors, corresponding to the five HBM components. These five sub-scales had reasonable reliabilities ($C\alpha$ = .70 -.84). In addition, hierarchical regression analysis indicated that readiness to seek counseling was predicted by: secure attachment (β = .12), perceived obstacles (β = -.09), perceived effectiveness of the intervention (β = .74), and by the interaction of avoidant attachment with the perception of problem severity (β = .13) and with perception of problem solving benefit (β = .09). Implications for career planning and counseling are discussed, especially those regarding career planning beliefs modification.

Clinical Assessment of Autism Spectrum Disorder: from Reality to Best Practices

Melanie Bolduc, Nathalie Poirier and Nadia Abouzeid
Université du Québec à Montreal, Montreal, Canada
e-mail: bolduc.mel@gmail.com

Autism Spectrum Disorder (ASD) affects 1% of the population and has a major impact on the lives of those who are directly or indirectly touched by it. Here we provide an in-depth analysis of the clinical assessment process, which evaluates the significant events that leads to a diagnosis of ASD. In our study, parents (N=50) of autistic child completed a self-administered questionnaire and performed an individual interview concerning several aspects of their children's condition. Furthermore, to fully grasp the scope of our analysis, our results were then compared with the recent guideline of best practices for the clinical evaluation of ASD, as recommended by the Quebec medical and psychological professional associations (Collège des médecins & Ordre des psychologues du Québec, 2012). Our results show a clear discrepancy

between actual clinical assessments of ASD and best practices recommendations. Indeed, while the guideline stipulates that health professionals should undertake screening of ASD early signs – e.g., developmental delays in language production – our results show that parents are usually the first ones to notice the first signs. Thereafter, our results also indicate that the delay to receive an ASD diagnosis - i.e. average of over 4 years from early signs detection - is divergent from the recommended best practices. Finally, we show that, as recommended, the diagnosis is generally based on behavioral observations, interviews and standardised evaluations performed by a multidisciplinary team. In conclusion, in line with the promotion of best practices in the detection, diagnosis and intervention, it is critical to address parents' concerns sooner, which entails faster diagnosis and earlier interventions.

Student Course Evaluation Is Revisited in Virtual Context

Bengu Borkan

Bogazici University, Istanbul, Turkey

e-mail: bengu.borkan@boun.edu.tr

While student's evaluations of course and teaching effectiveness have commonly used in North American academia, few colleges have been using it in developing countries like Turkey. Although literature and discussion in academia display support and opposition to use of students' course evaluation (SCE) to measure teaching effectiveness, it has been used widely and more universities and colleges are about to initiate SCE. Even though it has been using for variety of purposes such as promotions, tenure, merit reviews, and promotions, course selection guide for students, it has been only used for feedback purposes in Turkish public universities. This limited uses brings some challenges in the implementation of this tool in Turkey. The first focus of this study is to present the virtual SCE practices and experiences of one medium scale public university [15 thousand (graduate and undergraduate) students] in a large metropolitan city of Turkey. Course evaluation has been implemented in more than two decades and the system has been switched to virtual environment 4.5 years ago in this university. In the SCE literature it was shown that various factors (class size and the student expected grade are most emphasized) create bias results in course evaluation. Therefore the second focus of the study is to detect possible biases in SCE. Longitudinal data from more than 800 courses in each semester during 9 semesters are analyzed to examine the factors related with course characteristics (class size, electivity, grade level, and subject area) effecting student's participation to course evaluation and overall students' ratings for particular course and instructor. In this study these factors are reanalyzed, as well as with students' assigned course grade. Moreover, which dimensions of the effective teaching (as defined in the scale) are most affected by sources of bias are examined on classroom level and student level.

Testing Non-Native Dutch Speakers with the Cattell-Horn-Carroll-Based Dutch Cognitive Ability Test

Annemie Bos [1], Tierens Marlies[1], Magez Walter[2] and Decaluwé Veerle[1]

[1]University College Thomas More, Antwerp, Belgium; [2]Coordination team Antwerp for Psychodiagnostics (CAP vzw), Brasschaat, Belgium

e-mail: annemie.bos@thomasmore.be

The Dutch Cognitive Ability Test (CoVaT -CHC) is a new CHC-based intelligence battery for children and adolescents in Flanders (Dutch speaking part of Belgium), and can be used for individual or group assessment. The main purpose of the test is to provide insight in general intelligence as well as specific individual cognitive strengths and weaknesses for the broad

cognitive abilities Fluid Intelligence, Crystallized Intelligence, Short-term Memory, Visual Processing and Processing Speed. During the construction of the test cultural fairness was taken into account. In addition, the CoVaT-CHC consists of both verbal and language-reduced subtests thereby making the test useful for non-native Dutch speakers. Previous research showed that cognitive abilities derived from the Cattell – Horn – Carroll taxonomy are often considered to be invariant across culturally and linguistically diverse populations. The main purpose of this study is to investigate cognitive strengths and weaknesses on the CoVaT-CHC in cultural and linguistically diverse children. More specific, this study compares group profiles of native and non-native Dutch speakers on the CoVaT-CHC. A representative sample of approximately 2000 children completed the CoVaT-CHC. Participants ranged from 10 years, 0 months to 13 years, 11 months. During four sequential lessons, participating children completed the assessment at school in groups of 5 – 25. Results & will be presented at the conference and will be discussed in the context of previous findings.

Examining Response Time Threshold Procedures for the Identification of Rapid-Guessing Behavior in Small Samples

Janine Buchholz[1] and Joseph Rios[2]

[1]German Institute for International Educational Research (DIPF), Frankfurt, Germany; [2]University of Massachusetts Amherst, MA, United States

e-mail: buchholz@dipf.de

Assessments that provide little consequences to the examinee (low-stakes assessment) threaten the validity of score-based inferences as test-taking motivation may serve as an extraneous factor unrelated to the construct(s) of interest. A proxy of examinee motivation can be obtained through response times from computer-based testing in order to identify rapid-guessing behavior and utilized as a basis for filtering invalid data from unmotivated examinees. However, only two previous studies have evaluated different methodologies for setting response time-based thresholds for defining rapid-guessing behavior. These studies are limited in two respects: 1) small sample contexts have been ignored and 2) findings have been based on single applied datasets, which limit the generalizability of results. This study aims to evaluate 13 rapid-guessing threshold procedures in a small sample size context via a real-data simulation. Data come from a low-stakes reading comprehension assessment administered to 3,557 examinees. The adequacy of each threshold procedure will be evaluated by manipulating sample size (50/100/250) and "true" percentage of rapid responders (5%/25%/50%). "True" thresholds were set by two trained raters who evaluated response time frequency distributions for each item. Results will be evaluated based on type-I-error, power, percent agreement in identifying unmotivated examinees, mean scores differences after filtering, and convergent validity. Preliminary results indicated very low type-I-error rates (<.9%) across all 13 procedures. However, great variation in power was observed between the methods with accuracy rates ranging from 71%-92%. Preliminary findings demonstrated that methods differ in their accuracy of identifying "true" rapid-guessing. Such a result has implications for: 1) identifying unmotivated examinees and 2) score-based inferences from data filtering. That is, if examinee motivation is inaccurately classified, analyses based on filtered data may also be inaccurate. The final paper will investigate this hypothesis by evaluating mean score differences and convergent validity across methods.

Differential Item Functioning of the Self-Esteem Test for Adolescents

Joaquin Caso Niebla[1], Carlos David Díaz López[1] and Luis Lizasoain Hernández[2]

[1]Universidad Autonoma de Baja California, Ensenada, Mexico

[2]Universi dad del País Vaso, Spain

e-mail: joaquincaso@gmail.com

The self-esteem test for adolescents (STA) is one of the instruments used in Mexico for the characterization of this construct in educational contexts. Data confirm the theoretical structure of this test, and refer high levels of internal consistency and acceptable percentages of explained variance (Caso et al., 2011). However, there is not enough research to determine if the test and the items have metric equivalence regardless the social, cultural, linguistic and instructional group (Elosua, 2003). Thus, this study was proposed in order to know the differential item functioning (DIF) in the STA regarding gender in a sample of Mexican students. Involved 9,303 students (4,725 women, 4,578 men) from 16 schools in middle school education from Mexico City, with ages ranging between 10 and 18 years (x=14.0). Students responded the STA with 21 items which are grouped into four factors: self-cognitions, competence cognitions, family relations and rage. To perform the analysis of DIF, Linacre's procedure and Winsteps program (Linacre, 2012) were used. According to the parameters established in the following, only the item number 5 exhibits DIF (DIF CONTRAST = 0.56), suggesting that this item is more likely to be answered negatively by women than by men, presenting a slight to moderate DIF according to various reference systems (Zwick, Thayer & Lewis , 1999). These findings add to the results of previous research, highlighting themselves evidence of content, criterion-related and construct validity, which is significant contribution to the measurement of affective domain in this country.

Validation and Psychometric Properties of the Brazilian Five-Item Mental Health Index (MHI-5)

Juliane Callegaro Borsa[1], Bruno Figueiredo Damásio[2] and Silvia Helena Koller[3]

[1]Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil; [2]Federal University of Rio de Janeiro, Rio de Janeiro, Brazil; [3]Federal University of Rio Grande do Sul, Porto Alegre, Brazil

e-mail: juliborsa@gmail.com

Depression and anxiety are two types of mental disorders related to huge losses of quality of life, both to the patient as well as to their relatives. Beyond the personal and relational impairment, the cost of depressive and anxiety disorders due to loss of productivity and use of health services allocates them as the top-five most costly diseases, including all disorders. In the psychological and psychiatric literature, several screening instruments have been developed to quickly identify these diseases. Due to its psychometric properties, including specificity and sensibility, the 5-item Mental Health Index (MHI-5) is one of the most accepted and used tool. This study aims to evaluate the psychometric properties of the Brazilian version of the MHI-5. Participants were 524 subjects (69.8% women), aged from 18 to 88 years old (M = 38.3; SD = 13.26), from 17 Brazilian states. The sample was randomly split in two halves. An exploratory factor analysis (EFA), was performed with the first half of the sample (n1 = 262) and a confirmatory factor analysis (CFA) was then conducted with the second half of the sample (n2 = 262) to cross-validate the obtained exploratory factor structure. The robust maximum likelihood extraction method, in a polychoric correlation matrix was used. Reliability was assessed using Alpha coefficient index, composite reliability and average variance extracted. Convergent validity was evaluated using the Subjective Happiness Scale (SHS) and the Satisfaction with Life Scale (SWLS). Discriminant validity was evaluated by employing the 12-item General Health Questionnaire (GHQ-12). Exploratory and confirmatory factor analysis supported a single-factor solution. Reliability was assessed using alpha reliability, average variance extracted and

composite reliability. Adequate indexes of convergent and discriminant validity were also achieved. The scale presented strong evidences of validity and seems appropriate to evaluate mental health on the Brazilian population.

Psychometric Properties of the Brazilian Version of the Positivity Scale (P-Scale)

Juliane Callegaro Borsa[1], Bruno Figueiredo Damásio[2], Daiane Silva De Souza[3], Silvia Helena Koller[3] and Gian Vittorio Caprara[4]

[1]Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil; [2]Federal University of Rio de Janeiro, 6, Brazil; [3]Federal University of Rio Grande do Sul, Porto Alegre, Brazil; [4]Sapienza University, Roma, Italy

e-mail: juliborsa@gmail.com

The positive features of individual functioning have gained increased attention over recent decades in accordance with a view of well-being as a state in which individuals fully realize their potentials, successfully manage their lives and contribute effectively to their communities. Positivity is a basic disposition conducive to a positive appraisal of life and experiences. Considering the potential benefits that positivity may have in people's lives, a new measure comprising eight items has been developed to directly assess the construct: the Positivity Scale (P-Scale). This study presents the adaptation process and the psychometric properties of the P-Scale in the Brazilian context. Participants were 730 subjects (65% women), aged from 17 to 70 years old (M = 31.0 years; SD = 11.43), from 21 Brazilian states. The sample was randomly split in half to cross-validate the results. With the first half of the sample (n1 = 365), an exploratory factor analysis (EFA) was conducted. With the second half of the sample (n2 = 365), a confirmatory factor analysis (CFA) assessed the fit of the model obtained through EFA. Furthermore, convergent validity and group differences were evaluated. The EFA and CFA presented adequate and reliable results for the expected one-dimensional model [? = .86; SRMR = .045; TLI = .99; CFI = .98; RMSEA = .058 (.033 - .083)]. Moderate correlations were found between the P-scale and mental-health, subjective happiness and life-satisfaction indexes. The levels of positivity presented a low positive correlation with age, educational level and financial income. Slightly significant differences were found for occupational status and marital status. Positivity appears to be more related to personal dispositions than to sociodemographic aspects. Our results suggest that the P-Scale is a reliable measure to evaluate the levels of positivity in Brazil.

Cognitive Abilities of Visually Impaired Brazilian Children: A Comparative Study

Carolina Rosa Campos and Tatiana Nakano

PUC-Campinas, Campinas, Brazil

e-mail: carolene_crc@hotmail.com

Considering the importance of studies with specific population and the scarcity of psychological instruments aimed at assessing the intelligence of visually impaired children in Brazil, the objective of this study was to carry out a comparison between the cognitive performance of visually impaired children against two groups of children with normal vision (one group making use of vision, and the other not) through three subtests (verbal, memory, and logic) based on the Cattel-Horn-Carroll (CHC) model. The subtests were conducted on 47 children between 8 and 12 years old: 14 of which were visually impaired (M=10.28 years old; S.D.=1.58, six of them were girls and eight boys), 17 children with regular vision (M=9.94 years old; S.D.=1.43; all girls) and 16 children that had blind-folded (M=9.37 years old; S.D.=0.99, seven girls and nine boys). The results indicate significant differences regarding completion time of all subtests, with better

performance presented by children with regular vision. With regards to the number of incorrect cards opened in the memory subtest, significant differences were also found with children with vision requiring fewer attempts compared to the other groups. In conclusion, this study indicates that, due to the differences found between tested groups, a specific instrument for the assessment of the cognitive abilities of visually impaired children in the Brazilian population shows indeed relevant. With a larger sample size, the psychometrics properties of this subtests could be investigated in future studies.

Factor Structure of the Chinese Mandarin Version of the ASCA in a Sample of Chinese Students with Intellectual Disability

Gary Canivez[1], Yi Ding[2] and Paul Mcdermott[3]

[1]Eastern Illinois University, Charleston, United States; [2]Fordham University, New York, United States; [3]University of Pennsylvania, Philadelphia, United States

e-mail: glcanivez@eiu.edu

In 2010 Canivez, Ding, Kuo, Guo, Yang, and McDermott reported on the development and psychometric evaluation of a Chinese Mandarin translation of the Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993) with examinations of differential item functioning, factor structure, comparative item prevalence, and raw score differences with the U.S. standardization sample. To date there has been no examination of the ASCA-Mandarin with a sample of children with intellectual disabilities. Exploratory factor analysis of the ASCA-Mandarin core syndrome raw scores using multiple factor extraction criteria, oblique and orthogonal rotation, and coefficients of congruence are reported for a sample of 246 Chinese children with Intellectual Disability/Mental Retardation. 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, parallel analysis, and theoretical consideration). Minimum average partials and standard error of scree indicated that only one factor should be extracted. Oblique rotation (Promax) of the two factors indicated the ADH, SAP, SAI, and OPD core syndromes were associated with the first factor (Overactivity [OVR]) while the DIF and AVO core syndromes were associated with the second factor (Underactivity [UNR]). The correlation between Factor 1 (OVR) and Factor 2 (UNR) based on the promax rotation was -.052, supporting the independence of the Overactivity and Underactivity dimensions and viability of an orthogonal solution found in other ASCA studies. Orthogonal (Varimax) rotation of the two factors was used for final solution and coefficients of congruence and salient variable similarity indexes tested the factorial invariance of the present factor structure results compared to a large sample of normal Chinese students (N=554) and the total ASCA standardization sample (N=1400). The factor structure was similar to that previously observed with other samples of Chinese, Canadian, U.S. Native Americans, and U.S. Hispanic/Latinos.

Expectation and Performance on Intelligence Tests: Investigation of Gender Differences

Erika Carvalho Voigt and Solange Muglia Wechsler

Pontificia Universidade Catolica de Campinas, Campinas, Brazil

e-mail: erikavoigt.ic@gmail.com

People´s expectation of their intellectual abilities tend to influence their performance in intelligence tests. In particular, women tend to be influenced by stereotypes about their abilities in different intellectual areas. The purpose of this research was to compare the expectations that young people have about themselves on their intellectual capacity with their results in intelligence tests. The sample consisted of 60 young (19 M , 41 F) aged 16 to 20, students in

public and private high schools (3rd degree) in the city of Campinas (São Paulo , Brazil). The instruments used were the Battery of Creative and Intellectuals Skills (BAICA), composed of the subtests of Verbal Comprehension (synonyms, antonyms, analogies), spatial reasoning, logical thinking and creativity. A questionnaire with 20 statements in likert type scale format (agree / disagree), has been produced for self-assessment in 4 areas: Logical, Verbal, Spatial and Creative. The instruments were administered in two sessions in their own classrooms. All the results in BAICA were analyzed to verify gender differences (t-Student's test). Subsequently the areas of personal questionnaire were compared with tests measuring the same skills, by Pearson correlation The results showed differences between genders, since women perceived themselves with more skills in the verbal area, while men perceived themselves with more skills in the spatial area. The study showed an effect of gender and type of school, since girls in public schools perceived themselves as more creative than boys. There were no significant relationships among the results of intelligence tests and self-perception skills. We conclude that both genders do not know theirs skills or talents, having a distorted view of the same, which can hinder a better career choice.

Revision of Two Dimensions of the Dimensional Clinical Personality Inventory (DCPI): Conscientiousness and Attention Seek

Lucas De Francisco Carvalho

University of São Francisco, Itatiba, Brazil

e-mail: lucas@labape.com.br

In Brazil, the scientific publications dealing with personality disorders (PD) and instruments for assessing PDs are scarce. Considering that, was developed in 2011 the Dimensional Clinical Personality Inventory (DCPI), based on axis II diagnostic criteria of DSM-IV-TR and also considering the clinical perspective (integrative evolutionary theory of Theodore Millon). At the present time the instrument is being revised according to other proposals, including the personality disorders proposal for DSM-5 (Chapter 3). This study aimed to review two dimensions of the DCPI, Conscientiousness and Attention Seek. 240 participants of both sexes, aging from 18 years old, answered the DCPI, the Brazilian version of the NEO-PI-R, and a sociodemographic questionnaire. The internal structure of each dimension were analyzed using the exploratory structural equation model (ESEM), the reliability indices of the factors were analyzed by alphas's Cronbach coefficient, and the relationships between the factors of the instruments was verified using Pearson's correlation. Regarding Conscientiousness dimension, it met an internal structure composed of 6 factors, and a model with adequate fit indices, showing an average internal consistency coefficient of .84. The relations between the dimensions of the DCPI and the dimensions and facets of the NEO-PI-R were consistent with expectations. Regarding the Attention Seek dimension, a structure composed of 4 factors were founded, demonstrating adequate fit indices and average internal consistency of .83. Also in this case, the relations with the dimensions and facets of the NEO-PI-R corroborated what was theoretically expected. In general, there is a properly functioning for the dimensions reviewed the IDCP.

Psychometrics Properties Verification of the Scale of Subjective Well-Being (SSWB) Using the Graded Response Model

Lucas De Francisco Carvalho[1], Cristian Zanon[2], Rodolfo Ambiel[2] and Carla Fernanda Ferreira-Rodrigues[2]

[1]University of São Francisco, Itatiba, Brazil; [2]University of São Francisco, São Paulo, Brazil

e-mail: lucas@labape.com.br

The positive psychology has gained place on the world overview, and one of the first constructs studied in Brazil was the subjective well-being. This study aimed to verify the psychometric properties of the Scale of Subjective Well-Being (SSWB), a self-report instrument composed by three dimensions, applying the Graded Response Model independently to each component of the scale. Participants were 182 undergraduates of both sexes aged between 18 and 57 years (M = 24.6 years). Principal component analysis showed three dimensions, and additional analysis suggest that the three factors of SSWB are unidimensional. IRT analysis revealed that only one response category did not work as expected for one factor. It was found that participants tended to endorse more easily the items of the positive affects, and the lowest average of theta was on negative affect. The implications of these findings regarding the psychometric quality are discussed.

Children's Subjective Well-being: Testing Different Measures in Brazilian Children

Jorge Castellá Sarriera, Livia Maria Bedin Tomasi, Daniel Abs and Miriam Raquel Wachholz Strelhow

Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

e-mail: jorgesarriera@gmail.com

Assessing children's subjective well-being is an important challenge considering that few researchers have investigated and tested different well-being measures on childhood. The objective of this study is to test different measures of subjective well-being in a sample of 2,130 Brazilian children aged from 9 to 13 years old (M = 11.02, SD = 1.03), 45% boys, students from public and private schools in five cities of the Rio Grande do Sul State. The instruments are the Student Life Satisfaction Scale (SLSS, Huebner, 1991), with 7 items answered in a 5-point scale ranging from strongly disagree to very much agree, the Brief Multidimensional Student Life Satisfaction Scale (BMSLSS, Seligson, Huebner & Valois, 2003), composed of 5 Items answered in a 11-point scale, ranging from completely dissatisfied to completely satisfied and the Personal Well-being Index-School Children (PWI-SC, Cummins & Lau, 2005) composed of 7 items answered in a 11-point scale, ranging from completely dissatisfied to completely satisfied. Reliability analysis and confirmatory factor analysis were carried out to verify the internal consistence and the scales factorial structure. Multigroup confirmatory factor analysis was also performed to verify the scales configural, metric and scalar invariance regarding children's age. Results show a satisfactory reliability of the PWI (α = .703) and BMSLSS (α = .646). Two items of the SLSS presented negative factor loadings and also increase the scale's alpha when deleted (α goes from .346 to .805). Good fit index of the confirmatory models of the scales were also found (CFI > .960, RSMEA < .08). The multigroup models showed metric invariance for PWI and only configural invariance for SLSS and BMSLSS. With these data in hand, researchers can advance in development studies with the analyzed scales, and also considering the children's age.

Optimising 360° Performance Assessment Using Rater Group Data

Sarah Chan and Rab Maciver
Saville Consulting UK Ltd, Surrey, United Kingdom
e-mail: sarah.chan@savilleconsulting.com

To effectively measure work performance using multiple raters, it is important to understand both the reliability of different constructs and the inter-rater convergence of different rater groups. While there is some literature on these two topics, there is less on the inter-rater convergence of different constructs for different rater groups. To investigate the inter-rater convergence of different rater groups on different constructs using Wave® Performance 360. Correlations were run between and within rater groups on Performance 360 assessments containing raters (N=14,355) from four groups: Self, Boss, Peer and Report. The two constructs that had the highest average correlations between all rater groups were Establishing Rapport .38 and Meeting Timescales .37, and the two that had the lowest average correlations were Inviting Feedback .12 and Adopting Practical Approaches .16. Between-group correlation were highest for Boss and Peer .36, then Peer and Report .35, and Boss and Report .27. Correlations between Self and others were lowest .19 (Boss) to .21 (Peer). Within-group correlations for Peer and Report were .45 and .44 respectively. Between-group or within-group correlations generally followed the same pattern as the pattern of correlations for constructs, except: o While Boss and Report had the lowest agreement between any two groups (excluding Self), they agreed more on constructs including Interacting with People and Taking Action. o Within-Peer agreement was higher than Within-Report agreement for Developing Expertise, but lower for Making Decisions and Team Working. The variance in inter-rater convergence in 360° performance assessments has shown to be a function of both the nature of construct and which two types of rater are being considered (e.g. peer-peer, peer-boss). When using performance assessments as a criterion measure, applying a set value (across constructs and rater groups) to correct for reliability may produce erroneous results.

Numbers, a Telling Story: The Importance of Context for Testing Data

Hillary Chan
The Graduate Management Admissions Council, Reston, VA, United States
e-mail: hchan@gmac.com

When the topic of data protection is brought to our attention, it usually means customer privacy, but how do we deal with the misuse of our publically available testing data? Even academic research has misconstrued publically available testing data to reach misleading or irrelevant conclusions. How do we fight the misuse of our data? We must be careful to report data, not simply for the process, but with the context as well. The testing industry collects and publishes data to not only provide annual numbers, but to assess the current testing population. To ensure the prevention of faulty inference of data, such as to a general population instead of the testing population, giving context to data is vital. In this presentation, an academic research study misusing publically available testing data will be critiqued. Although not all misuses can be prevented, the presenter will give examples of how one organization reports not just data, but stories using several methods. Testing companies often publish annual reports; including clear methodology and context is one way to prevent the data in these publications from being used incorrectly without consequence or challenge. In addition, ad hoc reports give support to context of data at a personalized level. Furthermore, interpretation given by the analyst through publications, webinars, and industry and academic articles also aids in preventing erroneous by third parties. The presentation will highlight the importance of thorough data collection and

reporting in context, which enables appropriate interpretation of data, as a best practice for global assessment.

Univariate and Multivariate Dependability of the CELPIP-G Writing Test: Effect of Scoring and Number of Raters and Tasks

Michelle Y. Chen[1] and Amery D. Wu[2]

[1]The University of British Columbia / Paragon Testing Enterprises, Vancouver, Canada; [2]The University of British Columbia, Vancouver, Canada

e-mail: michellec2004@gmail.com

Research has found that writing tests, which require extended and constructed responses from test takers, generally suffer from low score generalizability (Brennan, 2000). This is because many facets within the testing context (e.g., rating scale and raters), besides the targeted ability, can contribute to the score variance (Huot, 1990). Generalizability theory (G-theory) is a helpful measurement model that can be applied to investigate the relative effect of these facets on accuracy and dependability of writing scores. This study examined the impact of scoring algorithms and the effects of the number of raters and tasks on the dependability of writing scores in the Canadian English Language Proficiency Index Program-General (CELPIP-G) Test. The CELPIP-G Test is used to inform high-stakes criterion-based immigration and citizenship decisions for language proficiency. The CELPIP-G Writing Test consists of two constructed-response tasks: writing an email and responding to a survey question. Each task was rated by two independent raters in four domains on a 1-to-5 scale. Scores of 157 test takers who participated in the 2013 pilot test were analyzed by univariate and multivariate G-theory. Univariate G-theory analyses were conducted on task-level scores with EduG (SSREWG, 2006) and multivariate analyses were conducted on domain-level scores by mGenova (Brennan, 2001). The G-studies were a fully-crossed random-effects design (test taker × task × rating). The results showed that (1) dependability (Phi) coefficients for absolute decisions were high for CELPIP-G writing scores (>0.85); (2) most of the variance in writing scores was due to test takers; (3) to maximize the score dependability, it's more efficient to increase the number of tasks than to increase the number of ratings per task; and (4) a larger gain in the composite score dependability was achieved when more weight was given to the second task.

The Study of Cognitively Diagnostic Assessment Analysis of Line Symmetry from 5th Grade to 9th Grade Students of Taiwan

Cheng Chien-Ming

NAER, Taichung City, Taiwan

e-mail: cheng1234@hotmail.com.tw

Traditional test design can not reveal a student's mastery of skills information by the response patterns, and thus to help students and teachers a better understanding of the meaning represented by scores, and make more efficient Learning. Therefore, some scholars advocate integrating the cognitive science and psychometrics (psychometrics), and developing new diagnostic assessment methods to help achieve teaching objectives. This new Diagnostic evaluation method is called the Cognitive Diagnostic Assessment (CDA) .Cognitive Diagnostic Assessment focuses on the relationship between the latent knowledge structure and the response process of students. The study used line symmetry as test content. The two dimensions of framework are VanHiele four levels as vertical axis and Vinner's concept definition and property as horizontal axis. The test items are designed, reviewed by the framework. The test was administered by the 5th to 9th grade students of Taiwan. The study analyzed the response

data of students by the DINA model of cognitive diagnostic assessment, investigated items' analysis, and assessed students' ability. The results found that 5th to 8th grade students approximately reached level 2 of VanHiele - analysis and 9th grade students reached barely enough level 3 - abstraction. As for the students' ability difference, 9th grade students were significantly higher than 8th grade students, and 8th grade students were significantly higher than 5th to 7th grade students but there were no significant difference among 5th to 7th grade students.

Assessing the Holistic Person in a Diverse World: A Challenge for Psychologists

Carmen Chilina León[1] and María Carolina Berrios[2]

[1]Asesores de Desarrollo Integral y Universidad Católica Andrés Bello. Caracas, Venezuela; [2]Escuela de Psicología. Universidad Católica Andrés Bello. Caracas. Venezuela

e-mail: adinsc@gmail.com

During the last 70 years, psychological science typically has utilized tests within the context of research to understand and characterize human behavior and development. These efforts have lead to broad and specialized bodies of knowledge, including a number of theories, methods, and strategies that offers a deep but segmented comprehension about human behavior. That this knowledge remains unbundled, lacking unity and cultural flexibility has the potential to have adverse effects in developing countries that lack resources and display sociocultural characteristics that differ from those found in more developed countries (Vivas, León y Berríos, in press), thus limiting the development of traditional psychological assessment, theory, and research. But developing countries are able to contribute with the psychological science designing methods that are based on the bodies of psychological knowledge and their experiences, leading to more integrated theories that, in turn, promote a more holistic, accumulative and socio cultural understanding of human behavior, development and research. The aim of this paper is to present a theoretical framework, the Human Development Integrating Model, based on the views of Vygotsky (1978), Bronfenbrenner and Ceci (1994), Sociocultural Psychology (Valsiner & Rosa, 2007), and a research framework, the Holistic Research Method (Hurtado, 2012). These Venezuelan frameworks provide integrated models that has guided Leon's line of research designed to promote an understanding of the global person in his context, focusing on their children's strengths, weaknesses, threats, and opportunities within their families, schools, and communities, and has led to integrate practice, training, and research in their sociocultural reality.

Altruism Moderating Physical and Mental Health under Occupational Stress: the Perspectives from Genes to Behaviors

Yating Chuang, Xiaofei Xei and Huiyuan Jia

Department of Psychology, Peking University, Beijing, China

e-mail: yating15@gmail.com

Occupational stress is of increasing concern because it has become a serious problem which can impact on employees' psychological perception and physical experience. Research has proved altruism had positive influence on occupational stress. Besides, scientists have preliminary discovered several genes associated with altruism including DRD4 VNTR, COMT Val158Met, 5-HTTLPR, and AVPR1a. Hence, taken the altruistic genes as a new perspective, we attempted to examine the moderating factors of employees' occupational stress by exploring the effects of altruistic genes,personal traits and behaviors. Two hundred nurses were recruited from a

hospital in Chengdu (China). As for gene perspective, participants' saliva was collected to test candidate genes. As for behavioral perspective,an experience sampling method with a longitudinal survey was adopted to collect data in 10 working days during two weeks. Participants were required to complete four surveys every day (twice at work and twice at home) via cell phone. An anticipated result was to (a) prove the particular genetic markers associated with altruism and have positive effects on occupational stress; then (b) probe that both altruistic genes and helping behaviors may moderate negative effects of work pressure in moods, body feelings, job satisfaction and performance.

Utilization of Homework Websites Among College Students

Hatice Çiğdem Yavuz[1] and Burcin Tek[2]

[1]Ankara University, Ankara, Turkey; [2]The Embassy of the People's Republic of China in Ankara, Ankara, Turkey

e-mail: hcyavuz@ankara.edu.tr

Homework websites provide platforms for easy and quick access to information related to course assignments in the undergraduate (as well as graduate) education. Therefore, their utilization has been steadily increasing among college students day by day. The main aim of this paper is to determine the profile of these students and understand how such homework websites contribute to their academic performance/success and need for cognition. In our assessment, data was collected from 103 college students with a survey conducted by the authors and Need for Cognition (NFC) Scale. The data gathered shows that academic performance of college students might be explained by approximately 2% of their need for cognition. As stated in the analysis of data, relationship between homework website utilization and need for cognition of college students is inconsistent. These results support that there is an insignificant correlation on the utilization of homework websites between either college students who have lower and higher need for cognition, or college students who have lower and higher academic success.

Reanalyzing the DISMAS Test Data: Comparing IRT and CTT Based Estimates of the Error of Measurement

Hynek Cígler, Michal Jabůrek and Jan Širůček

Masaryk University, Faculty of Social Studies, Brno, Czech Republic

e-mail: hynek.cigler@mail.muni.cz

DISMAS test (Traspe & Skalkova, 2013) is an original Czech assessment tool for the diagnosis of the structure of math disability in children in grades 1–5. It consists of five main areas divided into 14 unidimensional subtests and it was developed and standardized using classical test theory. Authors verified validity. Subtests' correlations with WISC-III were low, but significant, differences between normal and clinical population were high and significant for all grades. Regardless the high validity, reliability seemed to be poor – Cronbach alpha ranged from .44 to .95 on subtest level, on the test level from .69 to .76 depending on the grade level. The aim of this paper is to show why. We questioned the adequacy of CTT and internal consistency as a method for estimating reliability in this case. That is why we applied simple Rasch model and Partial Credit Rasch model to reanalyze the data. Consequently, the latent trait estimates as well as their error of measurement were transformed back to the raw score scale. We compared IRT and CTT based errors of measurement on the same scale for all observed raw scores according to the distribution of the standardization sample. In most subtests the error of estimate based on CTT was smaller around the average score but bigger in the extremes, especially at the low end

of the scale (DISMAS is designed to test disability, so it has a high ceiling effect). We conclude that IRT analysis produces more accurate estimate for children with math disability, for whom the test is designed, so in this case Rasch model seems to be more appropriate method to estimate confidence intervals than CTT. Finally, we discuss other sources of relatively low reliability.

Equivalence of Internet and Paper-and-Pencil Testing of the Big Five Personality Traits and the Social Desirability Hypothesis

Yann Le Corff, Véronique Gingras and Mathieu Busque-Carrier

Université de Sherbrooke, Sherbrooke, Canada

e-mail: yann.le.corff@usherbrooke.ca

Studies on the equivalence of Internet and paper-and-pencil tests have shown mixed results (e.g. Aluja et al., 2007; Joubert & Kriek, 2009; Meade et al., 2007; Ployhart et al., 2003). One suggested explanation is that social desirability could be lower in Internet tests (Heerwegh, 2009; Joinson, 1999). In this regard, it was suggested that personality inventories are more sensitive to a change of format due to their more intimate content (Ployhart et al., 2003). However, the only study that used a repeated-measures design, which eliminates the possibility of intergroup effects, supported the equivalence of Internet and paper-and-pencil personality testing (Salgado & Moscoso, 2003). First, we wished to assess the equivalence of Internet and paper-and-pencil personality testing and second, we wished to see if social desirability is indeed lower in Internet testing. 407 undergraduate students completed both an Internet and a paper-and-pencil version of the IPLC (Le Corff, 2013), a French-Canadian measure of the Big Five (Goldberg, 1990), and of a French translation of a short version of the Marlowe-Crowne Social Desirability Scale (Cloutier, 1994), with a one to three week interval. The paper-and-pencil assessment was completed in class and the Internet assessment was completed at the time and place chosen by the participants. Paired-sample t-tests indicated statistically significant differences in mean scores ($p < 0.05$) on each of the Big Five traits. Cohen's d calculated using Morris and DeShon's (2002) formula for within-subjects, however showed effect sizes to be small in all cases (varying from $d = 0.02$ to $d = 0.26$). Finally, mean scores on the social desirability scale were identical ($d = 0.00$). Results suggest that Big Five personality testing is robust across both assessment methods. Meanwhile the social desirability hypothesis was not supported. Finally, the study's limitations are discussed.

French-Canadian Adaptation of Achenbach's Adult Self-Report

Yann Le Corff[1], Éric Yergeau[1], Karine Forget[1], Catherine Proulx-Bourque[1], Annie Roy-Charland[2], John Tivendell[3] and Annabel Levesque[4]

[1]Université de Sherbrooke, Sherbrooke, Canada; [2]Université Laurentienne, Sudbury, Canada; [3]Université de Moncton, Moncton, Canada; [4]Université de Saint-Boniface, Winnipeg, Canada

e-mail: yann.le.corff@usherbrooke.ca

The aim of this study was to create a French-Canadian adaption of Achenbach's Adult Self-Report (ASR), an objective self-report measure of behavioral and psychological problems in adults. A backward translation by committee was conducted following the method proposed by Valleyrand (1989). Then both the original English version and the French-Canadian version were administered, with a one to two weeks interval, to 251 bilingual university students from four Canadian provinces (Manitoba, New-Brunswick, Ontario and Quebec). Results indicated that the correlation coefficients between the corresponding syndrome scales of the two versions varied between 0.72 and 0.87, with a mean correlation of 0.80. Cohen's d calculated using Morris and

DeShon's (2002) formula for within-subjects designs indicated that differences between mean scores varied from near-zero (d = 0.01) to small (d = 0.19). Cronbach's alpha coefficients were similar across the two versions and to those reported in the ASR's manual. They varied from 0.51 to 0.94 for the syndrome scales of the French version. These results indicate that the French-Canadian adaptation of the ASR is psychometrically equivalent to the original English version.

Using Tablets for Drivers' Psychological Assessment in Brazil

Flavio Costa[1] and Igor Pinheiro[2]
[1]Vetor Editora, São Paulo, Brazil; [2]UFSC, Florianpolis, Brazil
e-mail: flavio@flarc.com.br

For over five decades psychological assessments are mandatory for applicants of driver's license in Brazil. During this period, while many quantitative and qualitative developments occurred in urban mobility and in the vehicles themselves, the methodological processes and the constructs assessed from the psychological evaluation have remained unchanged. The psychological processes that mediate the relationship between the human mind and the traffic environment, therefore, presents relevant room for improvement, as significant theoretical advances regarding phenomena directly related to driving performance have been found and technological developments have occurred. The aim of the study was to develop a digital application to be used for the psychological assessment of driver license applicants, compatible with new theories and social demands. A tablet application was designed to evaluate the psychological phenomena of attention, working memory, and reasoning, according to the specifications for obtaining a driver's license in Brazil. Some evidence of ecological validity was sought by using moving stimuli consistent with traffic reality. Among the main advantages observed by the use of the new application stand out the sharp consistency of the instructions applied with both audio and text, the absence of scoring and grading mistakes due to the computerized procedures, the reduced mental workload required by the psychologists that no longer need to execute repeated tasks, and the easy electronic storage of data for ongoing analysis and research. Experimental subjects reported adequate understanding of the tasks, greater interactivity with the electronic items than with pencil and paper ones, and also a more enjoyable and less anxiogenic overall experience. The application of overall psychological assessment via tablet encouraged new psychometric studies that broaden and diversify its validity and reliability.

Visual and Verbal Material from the WAIS-IV Causes a Retroactive Interference Effect on Modality Specific Recall for the WMS-IV in non-clinical participants

Simon Crowe
La Trobe University, Australia
e-mail: s.crowe@latrobe.edu.au

Neuropsychological assessment often involves the administration of a number of test instruments in a single session, yet knowledge of the way in which these tests might interact remains limited. This series of two studies examined the interference effects associated with two very commonly employed measures; the Wechsler Memory Scale –Fourth Edition (WMS-IV) and the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV), with a view to determining proactive and retroactive interference effects associated with these instruments. Ninety-two non-clinical participants in each study were assigned to a two (interference vs. no interference) by two (retroactive vs. proactive) between-subjects design. The administration order of the tests was counterbalanced (i.e., administration of the WAIS-IV prior to the WMS-IV, and the WAIS-IV

administered during the delay interval of the WMS-IV), with a view to investigating possible proactive and retroactive interference effects. In Study 1, only the verbal memory components of the WMS-IV using the flexible approach and interference material from the WAIS-IV were employed. In study 2, visuospatial target material was employed. In Study 1, the results indicated a significant retroactive interference effect when verbal material was administered during the delay between immediate and delayed recall. In Study 2, a significant retroactive interference effect in delayed visual recall was noted. When presented proactively, the interference material in each study did not negatively affect memory recall. The interference effect of material presented during the delay phase of memory testing significantly impacted upon participants' measured cognitive performance. When using flexible battery approaches, clinicians must be particularly cautious in interpreting results from material presented during any memory delay as this might indicate a spurious diagnosis of a memory disorder. These findings support previous observations that administering conceptually similar material during the delay phase of a memory test will negatively affect memory recall.

Robust Estimation of Latent Ability Using 4PM-Robust Estimation

Buyun Dai

Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou, China

e-mail: 99341655@qq.com

In educational tests, there are response disturbances such as random guessing, carelessness, transcription errors. To address those response disturbances, Birnbaum and Barton & Lord respectively proposed three parameter model (3PM) and four parameter model (4PM). The 3PM and 4PM, however, increase the model complexity and difficulty in estimating item parameters. 4PM-Robust approach is proposed to reduce the Biases in the case of response disturbances. A guessing parameter ( ) and a carelessness parameter ( ) are involved in this approach. 4PM-Robust approach is compared with 2PM-MLE, Biweight estimation and Huber estimation in three simulation cases: (1) random guessing without carelessness, (2) carelessness without random guessing, (3) simultaneous carelessness and random guessing. The various values of and are also considered. Three simulation suggest 4PM-Robust (=critical point, =critical point) as an effective method for robust estimation of latent trait in the presence of random guessing and carelessness error. 4PM-Robust could be regarded as an improvement of 2PM-MLE. 4PM-Robust estimation is generally close to Biweight estimation and superior to Huber estimation, but its calculation is simpler than the both.

Effects of Statistical Models and Items Difficulties on Making Trait-Level Inferences: A Simulation Study

Bruno Damásio[1], Wagner De Lara Machado[2] and Nelson Hauck-Filho[3]

[1]Federal University of Rio de Janeiro, Rio de Janeiro, Brazil; [2]Federal University of Rio Grande do Sul, Porto Alegre, Brazil; [3]São Francisco University, Itatiba, SP, Brazil

e-mail: brunofd.psi@gmail.com

Researchers dealing with the task of estimating locations of individuals on continuous latent variables may rely on several statistical models described in the literature. However, weighting costs and benefits of using one specific model over alternative models is not always an easy task. The aim of the present simulation study was to compare the performance of seven popular statistical models (Sum scores; Principal Component scores; Maximum Likelihood EFA; Minimum

Rank EFA; Rating Scale; Graded Response, and Weighted Least Squares Mean and Variance Adjusted CFA) in providing adequate latent trait estimates. Fifteen unidimensional databases were simulated, considering three items difficulties distribution conditions (1- items targeted at the sample mean of latent trait distribution; 2 - items targeted at the lower tail of latent trait distribution; and 3 - items targeted at the upper tail of latent trait distribution) × five sample sizes (N = 100, N = 200, N = 500, N = 1000 and N = 2000). For each database, 10 items representing a continuous latent variable were generated. Pearson correlation and determination coefficient (r^2) evaluated the magnitude of correspondence between latent trait estimates. The t test, one-way ANOVA and factorial ANOVA were employed to evaluate mean differences and effects of simulation condition, sample size and statistical model on shared variance with the true latent score. Models tended to provide more accurate estimates of true latent scores when using items targeted at the sample mean of the latent trait distribution. The Rating Scale model, Graded Response model, and Weighted Least Squares Mean and Variance adjusted CFA yielded the most reliable latent trait estimates, even when applied to items inadequate for the sample distribution of the latent variable. These findings have important implications concerning some popular methodological practices in Psychology and related areas.

Measuring Meaning in Life: An Empirical Comparison of Two Well-Known Measures

Bruno Damásio[1], Wagner De Lara Machado[2] and Nelson Hauck-Filho[3]

[1]Federal University of Rio de Janeiro, Rio de Janeiro, Brazil; [2]Federal University of Rio Grande do Sul, Porto Alegre, Brazil; [3]São Francisco University, Itatiba, SP, Brazil

e-mail: brunofd.psi@gmail.com

The psychometric measurement of meaning in life (MIL) has a long tradition in the Humanistic and Positive Psychology literature. Although there are several instruments to assess MIL, several of them fail in presenting adequate convergent and discriminant validity. Considering these well-known problems, the MLQ-Presence and SoMe-Meaningfulness scales were developed to correctly evaluate MIL without overlapping its content with other related constructs (e.g., happiness, life satisfaction, depressive mood, etc). Despite these efforts, we believe that these two measures (MLQ-Presence and SoMe-Meaningfulness) may address slightly different psychological features. This study aims to investigate the extent to which the MLQ-Presence and SoMe-Meaningfulness measure the same latent trait of MIL. Participants were 2,094 subjects (64.4% women), ranging in age from 18 to 91 years old (M = 33.73; SD = 14.78), sampled from 22 Brazilian states. Initially, confirmatory factor analyses (CFA) focused on testing concurrent models for the factor structure of MLQ-Presence and SoMe-Meaningfulness instruments alone. The rationale behind these analyses was to ensure that each instrument presented adequate psychometric properties. In order to test our main hypothesis, a bi-factor CFA estimated whether 1) the SoMe-Meaningfulness and MLQ-Presence items could be predicted by a general factor of MIL, and 2) the items 4 and 5 from the SoMe-Meaningfulness scale could also be predicted by a latent variable of spirituality, independently specified. Confirmatory factor analyses (CFA) supported the unidimensionality of MLQ-Presence scale, but suggested that two distinct latent variables explained the items of SoMe-Meaningfulness scale. We further specified a bi-factor model in order to clarify the dimensionality of the SoMe-Meaningfulness scale. Results revealed that two items of the SoMe-Meaningfulness instrument primarily address spirituality rather than meaning in life. Our results suggest that the SoMe-Meaningfulness scale is still a measure that overlaps with spirituality issues, at least in the Brazilian culture.

Dilemma in the Construction of Educational Instruments to the Brazilian Reality

Cristina Maria D'antona Bachert[1] and Solange Muglia Wechsler[2]
[1]PUC Campinas, Sorocaba - São Paulo, Brazil; [2]PUC Campinas, Campinas - São Paulo, Brazil
e-mail: cristinabachert@hotmail.com

Brazil needs to improve the quality of education offered in Public Schools (83.5%) and Private Schools (16.5%). The majority of people of lower socio-economic level study in Public Schools, which have low effectiveness. This pilot study aimed to identify how teachers define teaching style and recognize some personal characteristics of their way acting in classroom. Took part of this survey 38 teachers who works in Elementary and High School, 60.5% of whom work in Private School and 39.5% in a Public School. They answered a questionnaire with eight open questions about teaching styles and aspects related to teaching effectiveness such as teacher's personal characteristics and features of students and school environment. The data collected have been interpreted by content analysis methodology. The results indicate that teachers explain teaching styles by their performance in classroom (28.8%), teaching methods (27.7%) and teacher's personal characteristics (16.8%). Although 91.4% of Private School Teachers recognize that their teaching style affects on student learning, only 36.8% of this group associate teaching styles and efforts to engage the students and optimizing the teaching-learning process. In the Public School teaching style is characterized by distance between teacher and students (28.2%), difficulting and postponing the construction of an interactive and co-operative learning environment. The preliminary gathered indicate that a Brazilian teaching styles assessment tool must contain everyday school situations that can bring future guidelines for teachers in order to enhance their teaching skills and improve the quality of education.

Acquiescence in Personality Questionnaires: Relevance, Stability, and Consistency

Daniel Danner and Beatrice Rammstedt
GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany
e-mail: daniel.danner@gesis.org

Acquiescence can be defined as agreeing regardless of the item content. Previous research has demonstrated that acquiescence can bias the correlation between items and the coherence of factor structures. We investigated the relevance, stability, and consistency of acquiescence in a big-five personality questionnaire. In particular, we investigated how much of the variance of manifest variables can be explained by acquiescence, how stable acquiescence is over time, and how consistent acquiescence is across different domains. Using structural equation modeling, we analyzed data from a representative German household sample (N=1286) and a heterogeneous Pannel study (N=160). The three core findings are (1) acquiescence in personality questionnaires accounts for about 5% variance of the manifest variables, (2) acquiescence is stable over time, and (3) acquiescence in personality questionnaires is only moderately related with acquiescence in attitude scales. These results suggest that we can easily improve measurement quality by controlling for acquiescence with structural equation models. Furthermore, the results suggest that acquiescence is a multidimensional, domain-specific construct which cannot be generalized from personality questionnaires to attitude scales.

Positive Hypothesis Testing as a Cause of Acquiescent Responding

Daniel Danner and Beatrice Rammstedt
GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany
e-mail: daniel.danner@gesis.org

Acquiescent responding, that is agreeing to an item regardless of its content, can bias the results of questionnaires. We investigated whether positive hypothesis testing is a cause of acquiescence in personality questionnaires, attitude scales, and knowledge questions. In general, answering an item can be described as a multistage process, where the respondent has to understand an item, collect information, make a judgment, and match the judgment to an answer category. We investigated whether hypothesis testing, that is collecting only information that can confirm an item, increases acquiescence. N=140 participants completed a total of 192 items and we measured agreement acquiescence (agreeing to all items) and acceptance acquiescence (agreeing to positively phrased items but not to negatively phrased items). Half of the participants were asked to consider only pro-item information. The other half of the participants were asked to consider pro- and contra-item information. In summary, the results from the positive hypothesis testing group demonstrate significantly greater agreement acquiescence in all domains and significantly greater acceptance acquiescence in attitude scales and knowledge questions. This suggests that positive hypothesis testing is one cause for acquiescent responding in questionnaires. From a conceptual point of view, we discuss relations between positive hypothesis testing and other potential causes of acquiescence such as cognitive resources and cultural background.

How to Enhance the Validity of Personality Assessments by Combining Big Five and Situation Perception in One Scale

Rudolf Debelak[1], Johanna Eisenhofer[2], Marco Vetter[1], Maria Pollai[1] and Matthias Ziegler[2]
[1]Schufried GmbH, Mödling, Austria; [2]Humboldt Universität zu Berlin, Berlin, Germany
e-mail: debelak@schuhfried.at

In the literature, a wealth of evidence for the validity of Big Five assessments in personnel selection and personnel development has been reported. Recently, it has been suggested to include situation perceptions into Big Five questionnaires because people's behavior does not only depend on tendencies as measured by the Big Five, but also on how they perceive and evaluate specific situations. This modification is suggested to further enhance the validity of Big Five questionnaires, however currently little is known on its actual effects. This study evaluates several aspects of validity of a new Big Five questionnaire which measures the Big Five personality facets and further provides an assessment of the individual judgment of specific business-related situations using a situational judgment test approach. Based on the data of 162 respondents (86 female; mean age = 42.21, std= 13.87), we first report results on the test's reliability and factorial structure, which confirm the expected structure of five personality and five situation perception scales. We then provide additional results on the relationship between the test results and the respondents' work engagement and income. These results indicate that the personality and situation perception scores provide more accurate predictions of these variables than the Big-Five scores alone. Overall, our study confirms that the validity of personality assessments can be enhanced by combining Big Five and situational perception items.

Assessing the Development of Bilingual Children from Turkish Immigrant Families: An Analysis of the Test Fairness of the Viennese Developmental Test

Pia Deimann, Ursula Kastner-Koller, Tugba Koc and Beyza Sahin

University of Vienna, Vienna, Austria

e-mail: pia.deimann@univie.ac.at

Preschool children from immigrant families have to cope with various developmental tasks. In particular, they are faced with the need for the concurrent acquisition of two languages, their mother tongue as well as the language of their host country. If bilingual development takes an adverse course, learning problems may result later on. Therefore, an early assessment of developmental problems due to bilingualism is crucial with regard to prevention. Developmental tests are usually standardized for monolingual children. A prerequisite for applying them to bilingual children consists in examining test fairness. The poster presents a study analyzing the test fairness of the Viennese Developmental Test Wiener Entwicklungstest, WET). The WET (Kastner-Koller & Deimann, 2002) is a developmental test for three- to six-year olds, which allows for a comprehensive developmental assessment of the entire range of important areas of functioning such as motor development, visual development, memory, cognitive and language development, psychosocial development. 30 Viennese children from Turkish immigrant families and 30 Turkish children of the same age, sex and socioeconomic background completed a Turkish version of the WET. Additionally the Viennese immigrants were tested with the German version of the WET. The language development of the bilingual Viennese children was retarded in both languages, while other areas of development were within normal range. The development of the Turkish children assessed by the Turkish version of the WET was consistent with the German standardization of the test.

AULA Virtual Reality Based Attention Test: Factorial Validity and Convergent Validity with EDAH Scale and DSM-IV Criteria

Unai Diaz-Orueta[1], Eduardo Garcia-Cueto[2], Beatriz Alonso-Sanchez[3], Nerea Crespo-Eguilaz[4], Manuel Antonio Fernandez-Fernandez[5], Cristina Otaduy[6], Carmen Perez-Lozano[7] and Aitziber Zulueta[8]

[1]Esplora, Technology & Behavior, Donostia-San Sebastian, Spain; [2]Universidad de Oviedo, Oviedo, Spain; [3]Centro de Psicología Bilbao (CPB), Bilbao, Spain; [4]Clínica Universitaria de Navarra, Pamplona, Navarra, Spain; [5]Institutp Andaluz de Neurología Pediátrica, Sevilla, Spain; [6]OtaduyVIP Centre, Valencia, Spain; [7]Proyecto 3 Psicólogos, Madrid, Spain; [8]ISEP Clinic Vitoria, Vitoria-Gasteiz, Spain

e-mail: udiaz@nesplora.com

AULA is a virtual reality based neuropsychological test for evaluating attention and support ADHD diagnosis in children between 6-16 years-old, with high test-retest reliability, sensitivity and specificity, and proven convergent validity with other attention tests. To study factorial validity of AULA and its convergent validity with EDAH scale and DSM-IV criteria. Two exploratory factorial analyses of the 18 main variables of AULA were performed with 2074 children from different Spanish schools and clinical centres. Both a 1-dimensional structure and a 3-dimensional structure (accounting for aspects of inattention, impulsivity and hyperactivity) were explored, by means of ULS extraction method. For the convergent validity with EDAH and DSM-IV, ADHD-subsamples of 188 and 360 children were respectively analyzed, performing cosine similarity analyses. Eighteen studied variables tend to saturate a single factor (F-values from .527 to .946), with 2 factors appearing as residual dimensions. The adequacy of the variables correlation matrix was analyzed in order to perform the factorial analysis (Bartlett = 55505.0, $p<.00001$; Kaiser-Meyer-Olkin = 0.89), indicating a good data adjustment (RMSEA = 0.071; GFI = .98; alpha = .98), with a total explained variance of 66% for the single dimension. In terms of

convergent validity, results show low-to-moderate correlations between AULA and EDAH, being the highest correlations for inattention (from .406 to .544). For DSM-IV, correlations are also low-to-moderate, being the highest correlation values from .379 to .473 for inattention. Results support the structure of AULA of one single factor that comprises the cognitive variables correlating with ADHD in any of its subtypes. With regards to convergent validity, different nature of AULA as an objective measure and EDAH and DSM-IV as observational scales suggest they target different aspects or dimensions of patients' behaviour and hence may complement each other in the increase of ADHD diagnosis accuracy.

A New Scale to Assess Pathological Video-Gaming among Adolescents Based on the DSM-V: An-IRT Based Analysis

Maria Anna Donati, Francesca Chiesi and Caterina Primi
University of Florence – Department of Neurofarba, Florence, Italy
e-mail: marianna_donati@yahoo.it

Video gaming is became very common among adolescents and high levels of game use frequently leads youth to pathological gaming. As a consequence, the assessment of pathological video-gaming is of increasing interest. Nevertheless, due to the lack of consensus about the definition of pathological gaming, there are some concerns in the field of measurement as many of the most widely used instruments were constructed by adapting pathological gambling criteria or using Internet addiction features. Indeed, only recently the American Psychiatric Association (APA) has introduced Internet Gaming Addiction in the Fifth Edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V, 2013). The aim of this study was to develop a new scale (Video-gaming Behavior Scale -VBS) to measure pathological gaming among adolescents. In constructing the items, we referred to the DSM-V diagnostic features for Internet Gaming Addiction, and we applied IRT to investigate the scale's accuracy along problem gaming severity levels. The validity of the VBS was also investigated taking into account the relationships between pathological gaming and risk seeking behaviors. The VBS was administered to 384 adolescents past year gamers (Male = 51%, Mean age=17.56, SD=.50). Once attested the one factor structure of the scale, undimensional IRT analyses for polytomous data were applied to evaluate the functioning of the VBS along the pathological gaming continuum. Item parameter estimates and the test information function showed that each item and the global scale satisfactorily measured the latent trait. Specifically, items had different levels of severity ranging from medium to high values and the test accurately measured medium-high gaming severity levels. Additionally, the positive relationship between pathological gaming and risk seeking behaviors provided evidence of the validity of the VBS. The VBS is an effective screening tool to identify adolescents who are pathological gamers.

A Rasch Analysis of the Sentence Completion Test for Youth

Mampaey Els and Caroline Andries
Vrije Universiteit Brussel, Brussels, Belgium
e-mail: els.mampaey@vub.ac.be

The Sentence Completion Test for Youth (SCT-Y)(Westenberg, Hauser & Cohn, 2004) measures the development of personal maturity of the respondent. The underlying theoretical framework for this test is the theory of ego development of Jane Loevinger (1976, 1985). Some studies, based on classic test theory, indicate satisfactory psychometric properties for the SCT-Y and give evidence of its construct validity (Westenberg et al., 2004). The objective of this study is to assess reliability and validate the Dutch version of the SCT-Y (Westenberg, Drewes, Siebelink,

Treffers, Jonckheer & Goedhart, 2000) using the Rasch model. The Dutch Version of the SCT-Y was presented to a sample of 1250 Flemish(1) children and young people between the ages of 8 and 25. Data were modeled using a constrained polytomous (rating scale) model (Andrich, 1978, 2010). Rasch analysis of the SCT-Y indicates good model fit, high test reliability and unidimensionality of the latent trait. There is proper item functioning for the majority of the testitems. Only one item has some drawbacks which make it essential to discuss arguments for exclusion or retention of this item. (1) The participants are inhabitants of Flanders, the Flemish speaking part of Belgium. Dutch is the language spoken in The Netherlands. All Flemish people can read and understand the Dutch test items.

Factor validity of the Polish version of Intelligence and Development Scales IDS

Diana Fecenec
Pracownia Testów Psychologicznych PTP, Warszawa, Poland
e-mail: d.fecenec@practest.com.pl

The presentation contains the results of the study on theoretical validity of the Polish version of Intelligence and Development Scales IDS — test used to assess cognitive, psychomotor, social and motivational development of children aged 5-10 years. The authors of the original version (Grob, Meyer, Hagmann-von Arx, 2009) assumed (and later empirically confirmed those assumptions) four-factor structure of the tool. The goal of the Polish study was to verify whether the Polish IDS adaptation retains the same four-factor structure. Exploratory factor analysis conducted with principal components method with varimax rotation was used to assess the factor validity. The analysis was based on IDS tests' standardized scores of children from Polish normalization sample (N=872). Four-factor solution was obtained confirming the theoretical validity of Polish version of IDS. However the same analysis applied separately to scores of pre-school kids (5-6 years, N=247) and school kids (7-10 years, N=535) has proven, that in the older children's group IDS has five-factor structure with perception being the fifth factor, giving ground to assumption thatthe start of school education has significant impact on children's cognitive functioning and leads to isolation of more specific cognitive abilities.

Assessment of Gender Differences in Strategies for Coping with Chronic Lumbar Pain

Sergio Fernando Zavarize and Solange Muglia Wechsler
Pontificia Universidade Catolica de Campinas, Mogi Mirim, Brazil
e-mail: sergio@fisiozavarize.com.br

Recent studies suggest that chronic pain manifests itself in a higher incidence and might be more intense for women. Although women usually live longer than men do, they use health services more frequently. The aim of this study was to investigate the gender differences on strategies to cope with chronic lumbar pain. The sample consisted of 158 participants, 105 women and 53 men, from 30 to 88 years old. Half the sample was aged 30 up to 54 years and the other half were from 55 to 88 years old. All of them presented chronic lumbar pain and were diagnosed with Lumbar Osteoarthritis. They were attending to five physiotherapy clinics specialized in pain treatment, in São Paulo, Brazil. The instruments used were the Visual Analogue Scale (VAS) which measures the pain perception; the instrument of World Health Organization Quality of Life (WHOQOL–BREF) which measures the quality of life; and a questionnaire elaborated to evaluate their leisure and distraction activities. The results demonstrated that women sought clinical treatment in greater number than men and also reported higher rates of pain - measured by the VAS scale. The Student t-test demonstrated a significant difference in pain perception between

genders (t = 2.657, p ≤ 0.01). Higher frequency of leisure activities and pain distraction were observed for women (22.76%). On conclusion, women have presented a greater number of pain coping strategies which probably have a higher positive influenced on their life quality. Therefore, women are more prepared to cope with chronic lumbar pain than men are.

On the Dimensionality of Group Development

Carlos Ferreira Peralta[1], Paulo Renato Lourenço[1], Paulo N. Lopes[1], Lucy L. Gilson[2] and Leonor Pais[1]

[1]Católica-Lisbon School of Business and Economics, Catholic University of Portugal and Faculty of Psychology and Education Sciences, University of Coimbra, Portugal, Lisboa, Portugal; [2]University of Connecticut, Connecticut, United States

e-mail: carlosferreiraperalta@gmail.com

Most models of group development predict that groups develop through four stages, along two subsystems: task and interpersonal. However, previous empirical research has been unable to consistently identify the eight dimensions of group development (four stages per subsystem) posited by these models. We believe a better measure is needed to assess these dimensions of group development. Our goal is to develop and evaluate a theory-driven measure to assess and identify the eight dimensions of group development. Data were collected from four samples: (a) 40 U.S. participants performing an item-sort task; (b) 396 U.S. group members rating items as key informants; (c) 102 Portuguese teams completing a questionnaire (based on results from samples a and b); and (d) 152 Portuguese group leaders completing a leader version of the questionnaire. Based on the item-sort task, 38 items were retained. Intraclass coefficients and average deviation indexes supported consensual validity, reflecting similarity of ratings across respondents within teams. Confirmatory factor analyses supported the eight-dimensional theoretical structure of group development in all samples used for that purpose (samples b, c and d). Reliabilities were always above .70. Measurement invariance revealed that key informants, teams, and leaders interpreted the meaning of the scale items in a similar manner. Correlation analyses supported the convergent and discriminant validity of group development in relation to theoretically related constructs such as trust in team members, turnover intentions, and satisfaction with team. We conclude that the 38-item measure is valid and reliable for the assessment of group development by key informants, team members, and leaders. Theoretically, we shed light on the dimensionality of group development and extend its nomological network. Practical implications for enhancing work-related outcomes via group development are discussed.

Examinee Motivation: A Global Challenge for Best Testing Practice

Sara Finney and Donna Sundre

James Madison University, Harrisonburg, United States

e-mail: finneysj@jmu.edu

The Standards for Educational and Psychological Testing specificies that testing programs should report examinee motivation to aid in test score interpretation. Similarly, the International Testing Commision calls for test users to "consider other qualities which may have artificially lowered or raised results when interpreting scores" (p. 15, Guideline 2.7.7; ITC, 2000). We investigated the influence of examinee motivation on the validity of test scores. First, we examined the properties of a test-taking motivation measure when there were different consequences associated with test performance (low-stakes for examinees vs. higher-stakes). Second, we investigated if perceived test importance, test-taking effort, and test performance differed as personal stakes of the test differed. During testing for institutional accountability

purposes (e.g., accreditation), students were randomly assigned to three different testing conditions where they were told one of the following: test scores have no personal stakes for students but scores do impact institutional reputation; scores have no personal stakes for students but students are sent their scores; scores are available for faculty review. We sampled incoming, first-year university students, in addition to students who had experienced at least two years of college. We supported the two-factor structure of examinee motivation (perceived test importance, test-taking effort) with both factors having adequate reliability across testing conditions. Moreover, we found negligible differences in perceived test importance, test-taking effort, and test performance across conditions and student populations. Fortunately, students gave similar and moderate amounts of effort during the test regardless if the test had lower or higher personal stakes. Given the number of tests that are low-stakes for examinees but are used for important national and international comparisons (e.g., TIMSS. NAEP), our results are encouraging. The validity of inferences from test scores is not automatically compromised in low-stakes testing contexts due to low test-taking motivation.

The Construct Equivalence of the Two Language Versions of the South African Substance Use Contextual Risk Index (SASUCRI)

Maria Florence[1], Susara Koch[2], Shazly Savahl[1], Serena Isaacs[1], Charl Davids[1] and Elron Fouten[3]

[1]University of the Western Cape, Bellville, South Africa; [2]Nelson Mandela Metropolitan University, Port Elizabeth, South Africa; [3]Rhodes University, Grahamstown, South Africa

e-mail: mflorence@uwc.ac.za

This paper forms part of a study that sought to develop and validate an instrument to measure perceived individual and contextual factors that are associated with adolescent substance use in low-socio economic status South African communities. The instrument is currently available in two languages, Afrikaans and English. The construct validity of the instrument was measured using procedures of construct and criterion-related validity. The aims of the current paper are to present findings on the construct equivalence across the two language versions of the SASUCRI. Cronbach's alphas for the two versions were compared using the following formula: $(1-\alpha_1)/(1-\alpha_2)$ which was introduced by van de Vijver and Leung (1997). The statistical significance of the group differences on all of the scales' reliabilities were used for this evaluation. The differences follow and F distribution with (n_1-1) and (n_2-1) degrees of freedom. The Tucker's Phi Coefficient of Congruence was used to assess congruence of the factor loadings across the two versions, at a scale level as well as at a second order factor level. No significant differences were found for the Cronbach's alphas for 14 of the 20 scales. The coefficients of congruence for 18 of the 20 scales as well as for one of the four second order factors were greater than .94 across the two versions of the instrument. This is an indication that evidence towards construct equivalence was not found at this stage. Further research at an item level needs to be conducted to identify problematic items to assess its effect on the construct equivalence of the instrument and help identify reasons for the current findings.

The Assessment of Anhedonia Traits in Non-Clinical Young Adults

Eduardo Fonseca-Pedrero[1], Mercedes Paino[2], Javier Ortuño-Sierra[1], Marta Santarén-Rosell[1], Serafín Lemos[2] and José Muñiz[2]

[1]University of La Rioja, Logroño, Spain; [2]University of Oviedo, Oviedo, Spain

e-mail: eduardo.fonseca.pedrero@gmail.com

Anhedonia has been hypothesized as a latent trait underlying risk for schizophrenia-spectrum disorders. The main goal of the present study was to analyze the psychometric properties of the

brief versions of the Physical Anhedonia Scale (PhyS-B) and the Revised Social Anhedonia Scale (RSAS-B). The final sample was comprised of a total of 1349 college students divided into two subsamples (n1= 710; n2= 639). Results showed that both brief instruments have adequate psychometric properties under Classical Test Theory and Item Response Theory frameworks. Internal structure analysis of PhyS-B and RSAS-B, through exploratory and confirmatory factor analysis, yielded an essentially one-dimensional solution. Cronbach's alpha coefficient for the total score of PhyS-B ranged between 0.86 and 0.87, whereas for the RSAS-B ranged between 0.88 and 0.94. Several items showed differential functioning by sex. The results indicated that the short version of the Physical and Social Anhedonia Scales showed adequate psychometric properties for the assessment of the negative schizotypy phenotype in this sample of young adults. Future studies should replicate the findings found in this research as well as conduct follow up studies.

The Structure of Maladaptive Personality Traits in Adolescence

Eduardo Fonseca-Pedrero[1], Mercedes Paino[2], Javier Ortuño-Sierra[1], Marta Santarén-Rosell[1], Serafín Lemos[2] and José Muñiz[2]

[1]University of La Rioja, Logroño, Spain; [2]University of Oviedo, Oviedo, Spain

e-mail: eduardo.fonseca.pedrero@gmail.com

The Personality Diagnostic Questionnaire-4+ (PDQ-4+) is a self-report used for the assessment of personality disorder traits, however, its psychometric characteristics have yet to be tested in community samples of adolescents. The main goal was to analyze the psychometric properties of the PDQ-4+ scores in a large sample of non-clinical adolescents (n = 1,443; M = 15.9 years; SD = 1.2). The PDQ-4+ scores showed adequate psychometric properties. Reliability of the subscales, incorporating a Likert-type 5-point response format, ranged from .62 to .85. The study of the internal structure at item level revealed that the PDQ-4+ subscales were essentially one-dimensional. Analysis of the internal structure at the subscale level by means of exploratory factor analysis and exploratory structural equation modeling yielded a possible three-dimensional solution. The PDQ-4+ subscales correlated moderately with emotional and behavioural variables measured by the Strengths and Difficulties Questionnaire. The results have clear implications for the understanding of maladaptive personality traits in adolescents.

A Measure of Cultural Resistance

María José García De La Barrera Trujillo and José Mafokozi Ndabishibije

Universidad Complutense de Madrid, Madrid, Spain

e-mail: toohtte@hotmail.com

This research aims to conduct a general examination of the situation of the level of cultural resistance of teachers, Spanish students and immigrant students, especially in high schools of the Autonomous Region of Madrid. To this end, some fundamental studies linked to sociology, multiculturalism and immigration are taken as a reference. The objectives of this study were to design and validate a set of questionnaires to measure the level of cultural resistance of teachers and students; and to figure out the average level of cultural resistance in the school context. A correlational study is developed: relevant variables such as beliefs or feelings as explanatory precursors of behaviors of both teachers and immigrant students as well as their native peers are measured. Data related to the academic context from a cultural perspective in addition to the overall academic performance is also included. However, this paper is focused mainly on teachers. As for the reliability indices of each questionnaire, Cronbach's alphas range from .81 for immigrant students, .91 for native students and .88 for teachers. Besides, on a 10

point scale the average level of teachers' cultural resistance in general amounts to 3.14 while the school's cultural resistance stands at 3.48: both average levels can be considered low. However, the cultural resistance level linked to politics seems to be much lower (2.56) than to the school context (4.23). Three questionnaires were successfully empirically tested and their indices of reliability computed after validation by experts. Furthermore, the average level of cultural resistance displayed by teachers can be considered low. They tend to be quite similar in most sociological characteristics. However, what seems to be a conclusive evidence is the link between teachers' achieved academic level and their cultural resistance level in the school context, especially with relation to the economic and political dimensions.

CTT Based Methods and IRT Based Methods of Score Equating

Mikel García and Paula Elosua
University of Basque Country, San Sebastian, Spain
e-mail: mikelgarciamarkina@gmail.com

Score equating is a fundamental process when working with different tests, since it represents the basic means available to guarantee an appropriate comparison of the scores. There are many situations in which test applications require different forms of the same test. For example, in an academic progress study, with the purpose of repeatedly measure the same individual or group. Also equating is commonly used in testing programs where tests forms are administrated to different examinee groups. In order to get score comparability scores need to be linked and equated. Equating can be accomplished using either Classical Test Theory (CTT) or Item Response Theory (IRT). The aim of this work was to compare CTT based methods and IRT based methods using empirical data coming from a language competence test. Two equating designs were used: equivalent groups design and non equivalent group design with anchor items.

Assessing Goodness of Fit in Item Response Theory Models

Mikel García and Paula Elosua
University of Basque Country, San Sebastian, Spain
e-mail: mikelgarciamarkina@gmail.com

Item response theory (IRT) is an approach based model to psychological and educational measurement that is being widely used due to its potential advantages over classical test theory (CTT) in solving some applied problems: test construction, test equating, differential item functioning detection, or item banks. However the usefulness of IRT models depends on the fitting between model and data and so, fitting model's data are a major concern when applying item response models to real test data. The aim of this work was to study systematically the fitting goodness of the 1PL, 2PL, 3PL, GRM and GPCM models using a computer simulated test data under a MonteCarlo simulation. Unidimensional and polytomous data were generated under the GPCM. In this work we studied: 1. The variation in item discrimination parameters (a), 2. The pseudochance-level parameters (c), 3. The test length, 4. The shape of the ability distribution. We assumed the most common use in test scores: Ranking order of the ability of the examinees based on the ability measured by the test. Since we used simulated data, true ability scores are known and served as the criterion against which the estimates of the ability score derived from the model and so, it could be judged using Spearman rank-difference correlations and the average discrepancy in ranks.

Developing Guidelines for a Multi-level Language Test

Rebeca García-Rueda
Universitat Autònoma de Barcelona, Bellaterra, Cerdanyola del Vallès, Spain
e-mail: rebeca.garcia@uab.es

Some Spanish regions have passed a new law that requires university students to certify Common European Framework level B1, or higher, in English in order to graduate. In the light of this legislation, all Catalan universities agreed in 2012 to develop a multilevel test, Certificat de Llengües de les Universitats de Catalunya (CLUC), which would evaluate and certify the candidate's competence in English by assessing writing, speaking, reading comprehension, listening comprehension and use of language. We propose a standardized protocol for the development of such high-stakes tests for assessing second languages. Building on Standards for Educational and Psychological Testing (AERA, APA y NCME, 1999), we aim to develop guidelines that are both more specific to language testing and more accessible to experts in the field. This protocol should assist test developers in such areas as: justifying the need for and purpose of tests, conceptual delimitation of the construct assessed, item creation and evaluation, statistical analysis of items, internal structure, estimation of reliability and evidence of validity and fairness. It should also be of use to test administrators and examiners in terms of selection, test administration and scoring, interpreting and communicating results. In order to develop these guidelines, representatives of all the stakeholders involved in test design will be consulted, i.e. experts in psychometrics, language testers, curriculum specialists, test developers, examiners, university representatives and test-takers. This presentation will detail the initial stages of the project.

Escape Decision-Making under Real Fire and Simulated Fire Conditions

Yang Gao and Hong Li
Tsinghua University, Beijing, China
e-mail: gaoyang2009012361@163.com

The present research aims to explore the effects of condition and memory on escape decision-making. We recruited mice rather than human-beings as participants, and conducted a preliminary training and two main studies to test our hypothesis. We used a 3 (condition: real fire, simulated fire, common) × 2 (memory: remembered and forgotten) mixed design. In study one, escape time was examined when only the familiar exit is available (exit1 or 2). In study two, exit choices were examined when the familiar and unfamiliar exits are both available (exit 1 and 2). The main findings are: (1) Escape time before forgetting under real fire condition is significantly shorter than the memory baseline, and also significantly shorter than that under simulated fire condition. (2) Escape time of mice under the three conditions is all significantly shorter than that of the control group. (3) When both the familiar and unfamiliar exits are opened and the familiar exit is in smoke, real fire group tends to choose the familiar exit, whereas the other two groups prefer to choose the unfamiliar exit. In conclusion, there is significant difference of the decision-making under real fire and simulated fire conditions. Mice under real fire conditions tend to adopt intuitive decision-making, whereas under simulated fire condition they do not prefer intuitive decision-making.

Development of Situational Judgment Test of Potential (SJTP) Using Partially Ipsative Measure

Xi Le Gao[1] and Gonggu Yan[2]

[1]Aceman Group consulting. co., Beijing, China; [2]Beijing Normal University, Beijing, China

e-mail: lerg@163.com

Potential is an important factor in talent selection and development. However, the definition and measurement of potential still lack clear standard and sound proof. We see potential as a inclination within an individual which will naturally emerge when he or she solve problems. From this point of view, ipsative measure and situational test may be proper ways to test potential. This article introduces the development of situational judgment test of potential (SJTP) based on partially ipsative measure. 4 dimensions of potential, namely cognition, emotion, adversity, and leadership potential are classified according to individual's inclination on thinking and acting, dealing with people and things. The test is consisted of 12 situations and 36 questions within them. Under each question there are 4 options which represent a potential category separately. Participants should allocate scores to the 4 options to reflect his/her agreement to every option and total scores can't exceed 8 under each question. The test-retest reliability of the 4 dimension is around .5. Several cases studies of validity are reported and advices on using SJTP are provided.

Relationships between WISC-IV Scores, Self-Perceived Ability and Self-Esteem

Sophie Geistlich, Sotta Kieng and Thierry Lecerf

University of Geneva, Geneva, Switzerland

e-mail: sophie.geistlich@unige.ch

Relationships between self-assessed intelligence and test performance are generally moderate (r = ± .30). Meta-analyses have indicated that gender plays an important role, because males provided higher self-estimates of spatial or mathematical abilities than females. It has also been shown that self-esteem influences self-estimates of intelligence. The objective of this study was to investigate the relationships between intelligence, self-perceived abilities (SPA) and self-esteem for girls and boys. To achieve this goal, the WISC-IV was administered. On the basis of the subtests scores, four standard scores (PSI, VCI, etc.) and five CHC composites scores (Gv, Gc, etc.) were used. Self-perceived of school abilities (language, mathematics, science) were obtained using a French adaptation of the Perceived Ability in School Scale. Self-esteem was assessed using 8 items of the MDI-C and a French adaptation of the Self-concept scale for children. These tests were administered to 174 non-clinical children aged from 7 to 12 years (95 girls, 79 boys). T-tests revealed that girls outperformed boys for the WISC-IV Processing Speed Index. No other significant differences were found for the WISC-IV. Regarding self-perceived school ability, no sex difference was found. No sex difference was found for self-esteem.Three hierarchical multiple regressions predicting self-perceived school abilities were performed. Contrary to expectations, self-esteem did not predict any self-perceived abilities. Processing Speed Index predict self-perceived mathematics, while Verbal Comprehension Index predict self-perceived science. In contrast with previous meta-analyses, our data did not support the hypothesis that boys gave significantly higher self-estimates than did girls. The present results did not support that boys tend to have higher self-esteem than girls. The present results could be due to the fact that young children were tested and because self-perceived of school abilities were tested rather than self-estimates of intelligence abilities.

Non-Equivalence of Subjective Well-Being Single-Item measures: Evidence from Chile

René Gempp[1] and Jose L. Saiz[2]

[1]Facultad de Economía, Universidad Diego Portales, Santiago, Chile; [2]Depto. de Psicologia, Universidad de la Frontera, Temuco, Chile

e-mail: rene.gempp@udp.cl

There is an increasing interest in studying subjective well-being (SWB). Although using different single-item measures of global SWB became a frequent practice, scarce evidence on their comparability is available. Lack of equivalence among measures jeopardizes the comparability of results drawn from different studies. The aim of this study was to examine the equivalence of five single-item indicators of SWB, using survey data from a nationally representative sample (n = 2532) of Chilean adults (PNUD, 2012). Three direct indicators ("life satisfaction", "happiness" and the "Cantril´s ladder"), and two reversed ones ("life dissatisfaction" and "life suffering") were included. To assess equivalence we tested whether single-items were parallel, tau-equivalents or congeneric measures of SWB, through confirmatory factor analyses. The best fitted model (□? = 36.101, df = 6, RMSEA =.04, TLI = .98, CFI =.98) reveals that only "life satisfaction" and "Cantril's ladder" are parallel measures and the best indicators of SWB (? = .79). To clarify the constructs reflected by these indicators, multiple regression analyses were performed using multiple-item measures of domain-specific satisfaction, affect frequency and recent event exposure, as predictors, and each of the five SWB single-item measures, as criterion. Different subsets of predictors relate to each indicator. "Life satisfaction" and "Cantril's ladder" are primarily explained by domain-specific satisfaction although the former is additionally explained by negative affects. "Happiness" is equally predicted by domain-specific satisfaction and positive and negative affects whereas "life suffering" is only predicted by unfavorable event exposure. No significant prediction was found for "life dissatisfaction". Only two indicators behaved as parallel measures of SWB. In addition, these indicators, as well as the others, were explained by different combinations of life satisfaction domains, affective states, and/or events exposure. Hence SWB single items failed to demonstrate adequate equivalence, at least in this population.

A Comparison of Item Response Theory, Confirmatory Factor Analysis and Binary Logistic Regression Methodologies for Exploring Measurement Invariance

Masoud Geramipour

Kharazmi university, Tehran, Iran

e-mail: mgramipour@yahoo.com

This study uses simulated data with known properties in different manipulated conditions to assess the power of confirmatory factor analysis, item response theory and Binary Logistic regression methods for assessing measurement invariance and true positive Differential Item Functioning. Results indicate that although neither approach is without flaw, the item response theory–based and Confirmatory Factor Analysis approaches seem to be better than Binary Logistic regression in most conditions.

Differential Item Functioning Analyses of the PISA Student Questionnaire across English and Turkish Versions

Semirhan Gökçe, Sevim Sevgi and Giray Berberoğlu
Middle East Technical University, Ankara, Turkey
e-mail: semirhan@gmail.com

Programme for International Student Assessment (PISA) provides information about educational settings across different languages and cultures, not only in reading literacy measures, but at the same time some student and school related constructs are taken into consideration. These are the constructs which are related to literacy skills, and also, should be seriously considered by education policy makers in order to revise or develop school curricula. In 2012, PISA focused on mathematical literacy and in the student questionnaire, besides family background characteristics, students' affective variables related to mathematics are also covered. Among the constructs included in the student questionnaire, (1) students' feelings about learning mathematics, (2) their confidence about having to do some mathematical tasks and (3) their feelings in problem solving experiences seem to be important variables to consider for a successful mathematics education. Since PISA collects data on multilingual versions of the scales, equivalency of the questionnaires across different languages and cultures is a prominent issue. In the present study, differential item functioning analysis (DIF) will be carried out for these three sub-dimensions for the purpose of evaluating translation fidelity and cultural relevance of the items used in the respective sub-dimensions across English and Turkish versions of the student questionnaire. In the study, the statistical equivalence of the items in the student questionnaire will be evaluated by IRT model, specifically Samejima's graded response model. This model is appropriate for polytomous items that are scored on Likert-type scale, which is ordinal in nature. All the analysis will be carried out using Multilog IRT computer program. The item parameters will be estimated using the marginal maximum likelihood method. The content of the items which are flagged as DIF will be evaluated in terms of translation fidelity and cultural relevance.

Factor Structure and Measurement Invariance of the Difficulties Emotion Regulation Scale (DERS) in Spanish Adolescents

Isabel Gómez Simón[1], Eva Penelo Werner[2] and Nuria De La Osa Chaparro[3]
[1]Parc de Salut Mar, Departament de Psicologia Clínica i de la Salut Universitat Autònoma de Barcelona, Barcelona, Spain; [2]Laboratori d'Estadística Aplicada, Departament de Psicobiologia i Metodologia de les Ciències de la Salut, Universitat Autònoma de Barcelona, Barcelona, Spain; [3]Unitat d'Epidemiologia i de Diagnòstic en Psicopatologia del Desenvolupament, Departament de Psicologia Clínica i de la Salut, Universitat Autònoma de Barcelona, Barcelona, Spain
e-mail: isabel.gomezs@e-campus.uab.cat

Emotion dysregulation is a unifying dimension of several psychopathological symptoms such as prolonged dysphoria, labile mood, high anger, persistent fear and excessive worry. Deficits in emotion regulation (ER) appear to be relevant to the development, maintenance, and promising treatment target in a broad range of mental disorders. The Difficulties in Emotion Regulation Scale (DERS) is the most comprehensive measure of emotion dysregulation to date, but the Spanish version has not been validated in adolescents. To provide evidence on factor structure and measurement invariance across sex, and internal consistency. A community sample of 642 (293 boys and 349 girls) Spanish adolescents aged 12-18 responded to the DERS. The original version of DERS is a 36-item self-report questionnaire measuring clinically relevant deficits of ER originally grouped into six subscales: Lack of emotion awareness, Lack of emotional Clarity,

Difficulties controlling impulsive behaviors when distressed, Difficulties engaging in goal-directed behaviors when distressed, Non-acceptance of negative emotional responses and Limited access to effective emotional regulation strategies. Confirmatory Factor Analyses revealed that our proposal to maintain the 28 items from the Spanish adult version but allocating them into the six original factors, as in previous adolescent samples, showed better fit than the other models analyzed (CFI = .932; RMSEA= .060). Weak (∆CFI = .006), strong (∆CFI = ?.010), and strict (∆CFI = .003) measurement invariance across sex was achieved. Internal consistency for the subscale scores was moderate to excellent (omega .69 to .90). We found some sex differences on subscale scores, but effect sizes were small (Cohen´s d < .20). We hope to provide promising data about the feasibility of using the instrument with adolescents in Spain and thus contribute to a better understanding of the ER construct and its relationship with different mental disorders.

Psychological Evaluation on the Context of Public and Private Safety in Brazil

Fernanda Gonçalves Da Silva, Marcela Reis and Cristiane Faiad

Universidade Salgado de Oliveira, Rio de Janeiro, Brazil

e-mail: fernandagoncalves.fgs@gmail.com

The psychological evaluation to enter in the safety area in Brazil is a requirement patterned by laws that are present since the judicial aspect until the psychologist action. This regulation is due to the type of job done by the public and private safety agents, which concern the use and carrying of firearms. Thus, to enter and remain on this kind of service, it is necessary to make psychological evaluations in which instruments can be used to show psychometric quality. The results need to be in agreement to the context in which they will be hired. In these terms, it was created the research group in psychological evaluation in public safety – APSP, proposing to facilitate discussions with approach in selection procedures, in which professionals carry firearms. The current work presents the construction and validation of a scale to measure the emotional instability with specific focus on public safety area (IESP), where the difference is the use of validation sample formed, exclusively, by police officers and the construction of a verbal and non-verbal conscientiousness for private safety (ECVNV) validated for this specific context and that brings up, as the main contribution, the possibility to evaluate people with difficulties for reading. The instruments have been reasoned on the theory of the five big factors and cover the studies on search of evidences of validation based on contents as well as in the internal structure of scales, considering, also, the pursuit of validation evidences based on external variables. The scales present validation evidences and open the way to new researches with focus on an interface of the psychological science and the public and private safety area.

The Use of Confirmatory Multidimensional Scaling to Assess Cross-Cultural Measurement Equivalence

Dalray Gradidge

Nelson Mandela Metropolitan University,Port Elizabeth, South Africa

e-mail: dalray.gradidge@nmmu.ac.za

Psychological wellness literature has reported on the evolution of wellness assessment tools along with the growth of wellness theories during the past 40 years. However, wellness assessment in South Africa has typically been conducted qualitatively or through the use of overseas measures. The Wellness Questionnaire for Higher Education (WQHE) was developed in response to the need for a South African wellness measure. While evidence of the metric equivalence of the measure exists, it is recognised that additional evidence is required to

empirically address the cross-cultural equivalence of the Western concept of wellness upon which this South African measure was developed. The overall aim of this study is to explore the cross-cultural equivalence of the WQHE to guide further refinement of it where necessary. This study is located within the field of wellness assessment in counselling across cultural groups in Higher Education, and is guided by the application of psychometric test theory as it relates to cross-cultural psychology. WQHE data was collected from a convenience sample of South African students of higher education in the Eastern Cape Province. A combination of both statistical and judgmental research methods was employed to investigate the structural and conceptual equivalence of the WQHE across three language groups. This poster paper will report on the use of confirmatory multidimensional scaling (CMDS) as a method to investigate measurement equivalence across cultures.

Comparing Traditional and Rasch Analyses of a ZKPQ Questionnaire Subscale on a Police Academy Candidates Sample

Mihaela Grigoras

Centre For Psychosociology, Bucharest, Romania

e-mail: mihagrigoras@gmail.com

This study examined whether Rasch analysis could provide more information than true score theory (TST) in determining the functioning and dimensionality of the Sensation seeking scale of ZKPQ. Subjects were 800 candidates to police academy completed ZKPQ questionnaire as part of a psychological evaluation. TST and Rasch analyses examining functioning and dimensionality of the Sensation seeking scale were performed. TST indicated that the subscale performed well and could be improved slightly by deleting one item. The subsequent Rasch analysis indicated that three worst fitting items. Deleting them improved validity without loss of reliability. Factor analysis using raw scores and principal component analysis of residuals for the measure were also performed. The study supports the use of Rasch analysis over TST reducing the number of items while maintaining reliability and improving validity.

Validation of the Index of Attitudes toward Homosexuals in a Caribbean Sample

Jill Gromer[1], Mike Campbell[2] and Donna-Maria Maynard[2]

[1]Florida State University, Tallahassee, FL, United States; [2]University of the West Indies--Cave Hill, Bridgetown, Barbados

e-mail: jillgromer@gmail.com

The subject of sexual diversity remains controversial in Barbados and throughout much of the English-speaking Caribbean. Research on homophobia in the region is needed, but existing measures have not been validated using Caribbean samples. The Index of Attitudes toward Homosexuals (IAH; Hudson & Ricketts, 1980) is one such measure of homophobia; it is a 25-item self-report instrument that uses a 5-point Likert-type response format. It was designed as a unidimensional measure. The psychometric properties of the IAH have been established in the United States (Siebert, Chonody, Rutledge and Killian, 2009) and Australia (Pain & Disney, 1996). The aim of the present study is to assess the psychometric properties and factor structure of the IAH in a sample of university students in Barbados. Participants (n = 428, 75% women, 73% of Barbadian nationality, mean age = 23.6, SD = 7.4) were administered the IAH, a demographic questionnaire, and the Sexual Prejudice Scale (SPS, Chonody, 2009). The SPS is a measure of antigay bias. Reliability analyses were conducted, as well as an exploratory factor analysis (EFA) using maximum likelihood estimation and oblique rotation. Missing data (< 2%,

MAR) were imputed using maximum likelihood estimation. The IAH was internally consistent (.93) and was strongly correlated with both subscales of the SPS ($r = .83$, $p < .01$; $r = .72$, $p < .01$), providing evidence of concurrent validity. The EFA revealed a factor structure similar to that reported in US and Australian samples, suggesting that the construct of homophobia may be similar in these nations. The IAH has sound psychometric properties. The results support the continued research use of the IAH in the English-speaking Caribbean.

The Effectiveness of Lipp Stress Control Training in Post-Menopaused Women

Claudiane Aparecida Guimarães[1], Ana Paula Justo[1], Rogério Henrique Cogo De Oliveira[2], Éder Félix Dos Santos[2], Louis Lipp[2] and Marilda Emmanuel Novaes Lipp[2]

[1]PUC Campinas, Uberlândia, Brazil; [2]Instituto de Psicologia e Controle do Stress, Uberlândia, Brazil

e-mail: claudianeguimaraes@yahoo.com.br

Cardiovascular disease is the leading cause of death in women.Considering the multifactorial aspect of this disease, in order to provide effective care it is important to have the contribution of various areas of expertise. Within the realm of the psychological area, it is the aim of Lipp Stress Control Training (TCS) to identify and the modify life habits and potentially harmful behaviors. The TCS is comprised by four pillars: anti-stress nutrition, relaxation, physical activity and cognitive restructuring. The aim of this study was to test the effectiveness of the TCS inpost-menopaused women, aiming at reducing cardiac risk by identifying and modifying the level of emotional stress and psychological traits associated with cardiovascular diseases. The sample was composed by 10 post-menopaused women, with an age average of 67.6 years. All were evaluated in clinical and psychological terms before and after the TCS in group, which lasted 8 weeks. Laboratory tests were performed (glycemia, triglycerides, cholesterol, and cortisol) before and after the TCS. Instruments used were: Lipp'sAdult Stress Symptoms Inventory; Lipp and Rochas´s Quality of Life Inventory and Spielberger State and Trait Anger Inventory. Pre-and post-training assessments indicated significant differences in stress level (decrease) and scores of anger trait, anger-in (reduction), andanger-out (increase). Improvement in quality of life in social, affective and health areas were observed. Analyses of the hemodynamic indicators revealed significant reduction of triglyceride in the post-treatment evaluation. Given the results obtained, it is possible to suggest that the TCS seems to be aneffective measure for the reduction of some of the cardiovascular risk factors, but studies with larger samples are necessary to assess the extent of its effectiveness.

The Evaluation of Item Selection Methods in Cat with Respect to Different Item Pool and Ability Distribution Parameters

Melek Gülşah Eroğlu[1], Nagihan Boztunç Öztürk[2] and Hülya Kelecioğlu[2]

[1]Gazi University, Ankara, Turkey; [2]Hacettepe University, Ankara, Turkey

e-mail: gulsah_eroglu@yahoo.com

In this study, it is investigated that how different item selection methods resulted in terms of measurement precision with changes in item pool and individual characteristics. The data was produced by simulation method. Maksimum Fisher Information (MFI), Efficiency Balanced Information (EBI) and a-Stratification (a-Str) were used among different item selection methods. In the study, two item pool with different c parameters [uniform (0.05;0.1) and (0.11;0.2)] and 3 different ability distribution (right skewed, normal, left skewed) were used. Starting rule was set as b=0 and stopping rule was set as 30 items. The results were presented by obtaining RMSE, bias, fidelity and aad values. Accordingly, for the MFI and a-Str item selection methods, the highest measurement precision was achieved in normal distribution and lowest

chance success group whereas the lowest measurement precision was achieved in negative skewed distribution and highest chance success group. For the EBI item selection method, the highest measurement precision was achieved in normal distribution and lowest chance success group whereas the lowest measurement precision was achieved in low skilled individuals and low chance success group. Additionally, for the negative skewed distribution group, the EBI method with the use of high chance success item pool and MFI method with the use of low chance success item pool delivered lower RMSE values. In normal distribution, lower RMSE values were obtained through MFI and EBI methods in both high and low chance parameter item pool. In positively skewed distribution group, lower RMSE values were obtained through MFI and EBI methods in both high and low chance parameter item pool As a conclusion, it can be proposed that MFI and EBI item selection methods produce better results in comparison to a-STR method with the conditions valid in the study.

The Investigation of the b Parameter Distribution and Sample Size on the Performance of Characteristic Methods

Eren Halil Ozberk[1], Akif Avcu[2] and Hülya Kelecioğlu[1]

[1]Hacettepe University, Ankara, Turkey; [2]Marmara University, Istanbul, Turkey

e-mail: erenozberk@gmail.com

The aim of the current study was to investigate the performance of two commonly used characteristic curve equating methods (Stocking Lord and Haebara) with varying b parameter distributions and sample sizes. Population parameters were generated with varying b parameter distributions (skewness_b. = -0.5/0/0.5). The descriptive statistics of the population parameters for Base Test X_a=1, sd=0.2; X_b=0, sd=1; b skewness= - 0.5/0/0.5; X_c=0.2, sd=0.02; Xtheta=0, sd=1. For Equated Test X_a=1, sd=0.2; X_b= -1, sd=1; b skewness= - 0.5/0/0.5; X_c=0.2, sd=0.02; Xtheta= -1, sd=1. Additionally, the size of the sample was determined at 7 different levels (150, 250, 500, 750, 1000, 1500, 3000). Bu using those ability and theta vectors response matrixes was generated where 80 unique items and 20 common items were used for generating them. Finally, this response generation operation was repeated 100 times for each condition. The data generation was achieved by using irtoys package in r (R Core Team, 2013). The performance of the CCM's was tested by using RMSE. For the conditions where b parameter is not distributed normally, RMSE values are not decreased monotonously with increasing sample size. Additionally, for both methods, different b parameter distributions give different results for lower sample size but as the sample size increase, the results become similar. Finally, even the differences were ignorable, Haebara Method yielded less error for b parameter for smaller sample sizes in all conditions. The results were discussed in the context of the relevant literature.

Confirmatory Factor Analysis of the Indonesian Version of the WAIS-IV

Magdalena Halim[1], Christiany Suwartono[2], Lidia Hidajat[2], Marc Hendriks[3] and Roy Kessels[3]

[1]Faculty of Psychology, Atmajaya Catholic University of Indonesia, Jakarta Selatan, Indonesia; [2]Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia; [3]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Netherlands

e-mail: magdalena.halim@atmajaya.ac.id

The Indonesian version of WAIS IV has just been developed. The primary result from our first pilot study (Suwartono et al, 2013) seemed to be psychometrically promising. Therefore on this study we examined the factorial structure of Indonesian translated version of WAIS-IV in Indonesian sample using first order then second order confirmatory factor analysis. This study tested hypothesis regarding the 15 subtests of WAIS-IV can be conceptualized with four factor

then one general intelligence at the end. Using convenience sampling method, we had 444 participants (Men = 41%, Women = 59%), aged 16 – 69 years of age (M=33, SD=13.57), and had various educational background, and ethnicities from several cities in Java island. Results proved the four factor structure was supported by first order analysis and then confirmed with second order of confirmatory factor analysis. It was superior in fit to one factor structure. This result indicated the same four factor structure (verbal comprehension, perceptual reasoning, working memory, and processing speed) that was evidenced in US sample was evidenced in Indonesian sample. Even though, it was concluded that the WAIS–IV structure in Indonesia is the same as US, we also found that 15 subtests of WAIS-IV can directly fit into one factor, the general intelligence. These results lead to the discussion about the application and clinical use of WAIS-IV in Indonesia.

Comparing the New Mplus Alignment Feature with Traditional Multiple-Group Confirmatory Factor Analysis Methods for PISA Indexes

Leslie Hawley[1], Carina McCormick[1], Betty-Jean Usher-Tate[2] and Sara Gonzalez[2]

[1]Nebraska Center for Research on Children, Youth, Families & Schools, Lincoln, United States; [2]Buros Center for Testing, Lincoln, United States

e-mail: lhawley2@unl.edu

Multiple-group factor analysis typically considers three degrees of measurement invariance—configural, metric, and scalar (Milsap, 2011). It is necessary to achieve scalar invariance in order to appropriately compare factor means and intercepts across groups (Brown, 2006). However, scalar invariance rarely fits data well, especially with large numbers of groups (Muthén & Asparouhov, 2013). For this reason, Asparouhov and Muthén (2013) developed a new method called alignment to better address comparisons with large numbers of groups. The alignment method uses a simplicity function that does not require exact measurement invariance to estimate group-specific factor means and variances. The purpose of this study is to demonstrate and evaluate this new method for evaluating invariance across groups in complex international survey data (i.e., PISA). Alignment results will be compared to traditional multiple-group confirmatory factor analyses (MGCFA), and the utility of the two methods across various types of scales will be discussed. Data for this project were obtained from the 2009 Programme for International Student Assessment (PISA) assessment. For the purpose of our study, we will evaluate three student completed scales: 1) approaches to learning, 2) teachers' stimulation of reading engagement, and 3) enjoyment of reading. Three scales were chosen because the scales vary in terms of items, types of constructs, and presence of subscales. Data will be analyzed in Mplus version 7.1 (Muthén & Muthén, 1998-2012). This proposal represents a work in-progress. Data have been acquired and the initial steps have begun for the analyses, so there is sufficient time to complete the analyses in advance of the summer conference. This illustrative example will allow us to evaluate the potential utility of the new alignment method as well as contribute to the substantive research regarding its application with different types of scales.

Investigating the Moderating Effects of School Accountability Policies on the Relationship between Teacher Practices and Student Outcomes Internationally

Leslie Hawley[1], Betty-Jean Usher-Tate[2], Carina McCormick[1] and Sara Gonzalez[2]

[1]Nebraska Center for Research on Children, Youth, Families & Schools, Lincoln, United States; [2]Buros Center for Testing, Lincoln, United States

e-mail: lhawley2@unl.edu

School climate can influence students' motivation to learn and achieve (Deal & Peterson 2009). An aspect of school climate that influences teachers and subsequently students are the evaluation practices. In particular, an emphasis on accountability (e.g., students' test performance) may lead teachers to fear the threat of consequences (Santiago & Benavides, 2009. Consequently, a climate focused on accountability may lead teachers to alter their teaching practices, potentially influencing students' motivation and achievement. The purpose of this study is to evaluate the degree to which schools' teacher accountability policies moderate the relationship between teacher practices and student outcomes (motivation and academic achievement). The 2009 Programme for International Student Assessment (PISA) will be used for this project. A three-level (i.e., L1: student; L2: school; L3: country) hierarchical model will be analyzed using the HPMIXED procedure in SAS version 9.3. We will conduct two analyses with separate dependent variables: 1) teachers' stimulation of reading habits scale and 2) reading performance assessment score (incorporating plausible values methodology). Both analyses will include measures of teacher practices as predictors and accountability practices as moderators, as well as additional student- and school-level predictors to account for the socio-economic context of students, schools, and countries, providing policy-relevant results. This proposal represents a work in-progress. Data have been acquired and the initial steps have begun for the analyses, so there is sufficient time to complete the analyses in advance of the summer conference. One of the benefits of using hierarchical modeling is the ability to separate between-group and within-group effects of evaluation policies across schools and countries. Consequently, this research will contribute to the substantive literature regarding evaluation policies with student achievement data across different organizational and cultural contexts and expand existing literature by comparing these relationships in many countries.

Measurement of Gains in Intelligence – An Example of a Study Design and a Method of Data Analysis

Anna Hawrot and Aleksandra Jasinska

Educational Research Institute, Warsaw, Poland

e-mail: a.hawrot@ibe.edu.pl

In order to be able to estimate absolute differences between two given measurements of any variable, it is necessary that they are expressed on a common scale. The easiest way to achieve that is to use the same test in subsequent measurements. However sometimes it is not possible, e.g. due to the fact that the test measures too limited range of the variable of interest. The aim of the paper is to present an example of a study design and a method of data analysis allowing to estimate students' intelligence gains in a situation similar to the aforementioned one. We used data collected during a longitudinal study carried out as part of the "Research project for the development of educational value-added models". Random sample of 150 Polish lower secondary schools (5249 students) took part in the study. Intelligence was measured twice: when students were in the first (Raven's Standard Progressive Matrices) and in the third grade (selected items form Raven's Standard and Advanced Progressive Matrices). The results of two measurements were equated with use of the nonequivalent groups with anchor test design

(NEAT) with internal anchor. The statistical model was modified to take account of administering the tests to the same sample of students. Differential Item Functioning of the anchor items was verified. 2PL IRT model was used to estimate students' intelligence levels. The model allowed us to equate the results of two intelligence measurements and to estimate intelligence gains between the first and the third grade. Mean gain in intelligence equals 0,6 of standard deviation of results in the first grade. Furthermore, the size of the gains differs between schools. The results prove utility of the chosen study design and method of data analysis in estimating intelligence gains among adolescents.

Proposing a Shortened Version of Scandura & Graen's (1984) Leader-Member-Exchange Scale: Information Function and Validity

Ana Hernandez Baeza and Vicente Gonzalez Roma
University of Valencia, Valencia, Spain
e-mail: Ana.Hernandez@uv.es

Leader-Member-Exchange (LMX) (Dansereau et al., 1975) is a prolific theory for organizational research. According to this theory leaders develop a differential exchange with each of their subordinates, and the quality of these LMXs influences subordinates' attitudes and behaviors at work. One of the most common instruments to measure LMX is Scandura and Graen's (1984) 7-item scale. Even if this scale is not large, researchers frequently need to maximize the efficiency of their efforts to gather data by collecting information about a large number of variables. This makes it necessary to use very short scales with similar psychometric properties than the original scales. The aim of the study was to propose a shortened version of Scandura & Graen's scale by choosing the items that keep most of the information of the original scale. Samejima's (1969) Graded Response Model (GRM) is fitted to obtain the corresponding information functions for the original and the shortened scale. In addition, the two versions are compared in terms of criterion validity, considering a number of relevant criteria (e.g. well-being, commitment and propensity to leave). We use a sample of 426 health center employees to fit the GRM to the scale and evaluate the information functions and concurrent validity of the two versions analyzed. Moreover, for 182 of the employees, we use longitudinal data to compare versions' predictive validity. The 7 items clearly differ in their information functions. From the results we propose a shortened version made of three items that keep most of the information along the latent continuum. Regarding concurrent and predictive validity, both the pattern and the significance of correlations are similar for the original and the shortened version. The shortened version of the LMX scale is a good alternative to the original one both in terms of the information provided and criterion validity.

Education Level and Gender Differences on Spanish WAIS-IV Performance

Ana Hernández Fernández, Frederique Vallar and Èrica Paradell Calbó
Pearson Clinical and Talent Assessment, Barcelona, Spain
e-mail: ana.hernandez2@pearson.com

Clinicians are frequently required to estimate an individual's premorbid functioning. Procedures for adjusting the Wechsler scales for demographic variables (e.g., educational level) may buffers the effects of clinical status in predicting premorbid intellectual and memory functioning. Objectives. This study examines gender and education level differences on WAIS-IV performance in a Spanish sample. A Spanish general population sample (N=1002) stratified by gender, education level, region and population density was assessed with the Spanish WAIS-IV (Wechsler, 2012). ANOVA analyses were conducted to examine performance differences

between four education level groups (without studies, first grade, second grade, and third grade); and t Student analyses were conducted to examine gender differences. ANOVA results between education level groups showed statistically significant differences on all subtests, except Visual Puzzles; higher performance was shown in higher education level groups. Males and females means showed statistically significant differences on 7 subtests; males had higher scores on Block Design, Digit Span, Arithmetic, Visual Puzzles, Information, Figure Weight and Picture Completion. Differences between education level groups and intellectual performance support results from previous research and the relevance of demographically adjusted norms to answer clinical questions about premorbid functioning.

Effects of the Presence of DIF on Assessment and Diagnosis Decisions across Different Cut-Off Points

Dolores Hidalgo[1], Francisca Galindo[2] and Juana Gómez-Benito[3]

[1]University of Murcia, Murcia, Spain; [2]Vrije Universiteit Medisch Centrum, Amsterdam, Netherlands; [3]Universitat de Barcelona, Barcelona, Spain

e-mail: mdhidalg@um.es

In clinical and educational contexts cut-off scores are commonly used in assessment and diagnosis. An important aspect to ensure the validity of these scores is to obtain evidence about measurement invariance across different groups. This implies differential item functioning evaluation. This work raises the question about the maximum number of items with DIF that can be tolerated in a test in a way that does not substantially affect decisions made with cut-off scores. Thus the aim of this study was to determine the consequences of the presence of items with DIF in the interpretations of scores based on cutoffs. A simulation study was designed. Different variables were manipulated: (a) sample size in the reference and focal groups (b) amount of DIF in the test and (c) percentage of DIF items in the test. Item responses were generated using the one-parameter logistic item response model. The effect of DIF in the selection rate was analyzed by evaluating the difference between proportions of subjects that exceed the cut-off for independent samples (reference group vs focal group). Ln Odds-Ratio (Ln-OR), Risk Ratio (RR) and d transformation for OR were calculated to assess the effect size of mean difference. The statistical test detected erroneously differences between the focal group and the reference group in the conditions of highest degree of contamination test. In large sample sizes and tests with a small percentage of items with DIF (10% of items) differences in the rate selection between groups were detected. The presence of DIF items in a test has an important effect on the based on a cut-off score.

Differential Item Functioning Detection Using Logistic Regression: A Systematic Review Literature

Dolores Hidalgo[1], Mª Dolores López-Martínez[1], Georgina Guilera[2] and Juana Gómez-Benito[2]

[1]University of Murcia, Murcia, Spain; [2]University of Barcelona, Barcelona, Spain

e-mail: mdhidalg@um.es

Systematic review of scientific literature is a useful methodological tool in order to obtain scientific evidence to support decisions in both applied and research context. Furthermore, Differential item functioning (DIF) is a key topic in relation to tests applied in educational, psychological, social, and health settings. Moreover, among the numerous statistical techniques that have been proposed to detect DIF, logistic regression is one of the most widely recommended because it provides a general framework for analyzing and evaluating DIF in relation to different aspects. The aim of the study was to present a systematic literature review

on the state of the art of the use of the Logistic Regression procedure for Differential Item Functioning detection. The search was conducted through ERIC, PsychInfo and Web of Science covering the period between 1990 and December 2013. Papers were included if Logistic Regression and Differential Item Functioning were mentioned in title or abstract. A total of 251 papers, coming from more than 80 different journals, were included in the analysis. The publications were simulated DIF papers or papers reporting applied DIF studies using this technique. The 88% of the papers were published from the XXI century, whereas that the 50% of the papers were published in the last five years. Systematic review literature studies can provide additional information that complements reviews about the functioning of different techniques for detecting DIF. By describing the studies carried out to date the results of this systematic review highlight several relevant issues for practitioners and researchers who use Logistic Regression technique as a way of detecting DIF.

The 3P Leadership Model: Validation Using 360 Data

Tom Hopton and Rab Maciver
Saville Consulting, Surrey, United Kingdom
e-mail: tom.hopton@savilleconsulting.com

We have developed a new, validated leadership model which goes beyond existing leadership theory (e.g. Judge et al., 2002; Vroom & Jago, 2007) to consider both situational and broader aspects of leadership. Our 3P model has been validated from conception and has its origins in established 360 and self-report behavioural data. This paper presents criterion-related validity evidence for the 3P model, based on 360 appraisal data. It also introduces the new concept of Pioneering Leadership. N=308 individuals completed Wave Professional Styles. For each of these 308 individuals a set of independent performance ratings was collected concurrently from a stakeholder. The 3P leadership scales were explicitly aligned a priori to performance criteria in order to clearly demonstrate their individual contribution to forecasting leadership effectiveness. The broad 3P leadership scales show stable and positive relationships across a range of different performance criteria. Professional Leaders are likely to be particularly effective at applying specialist expertise, evaluating problems and structuring tasks. People Leaders are likely to be particularly effective at accomplishing objectives, communicating information and providing leadership to a wide range of people. Pioneering Leaders are likely to be particularly effective at demonstrating potential, creating innovation and driving success. The results provide validity evidence supporting our 3P model. The results suggest that while there are different types of leaders which can be effective in different situations, the most effective leaders tend to score high on the primary scales of the 3P model.

The Influence of Self-esteem, Activity and Mood on the Implicit Self-Appraisal

Dmitry Inozemtsev
South Ural State University, Chelyabinsk, Russia
e-mail: subonair@gmail.com

The aim of this study is the influence of Self-esteem, Activity and Mood on the implicit self-appraisal. Sample consisted of Russian students (N = 90). Measurement the implicit association test, questionnaire "mood", an experimental procedure based on the Raven‘ s Progressive Matrices The experiment consisted of three phases. Implicit self-appraisal and mood were measured at the first stage. In the second stage people took part in experimental procedure based on the Raven’s Progressive Matrices. Results were deliberately understated. In the third stage implicit self-appraisal and mood were measured at the second time. We divided the people

into two groups, depending on their results on the questionnaire "mood". Experimental group consisted of people who have a difference in values between the first and the second measurements of the questionnaire. The second time results measurement of IAT in the control and experimental group were also compared between ourselves. Level of implicit self-esteem in the experimental group significantly different from the level of implicit self-esteem in the control group ($U = 330$, $p < .01$). The study results suggest that implicit self-esteem varies depending on whether there is a change in the mood.

A Framework to Review Research in TIMSS Turkish Sample

Sevgi Ipekçioglu[1], Serkan Arikan[2] and Semirhan Gökçe[1]

[1]Middle East Technical University, Ankara, Turkey; [2]Muğla Sıtkı Koçman University, Muğla, Turkey

e-mail: isevgi@metu.edu.tr

TIMSS has been a longitudinal comparative program which aims to measure mathematics and science achievement of participating countries at the fourth and eighth grades on every 4 years since 1995. TIMSS results can be effective in gaining a deeper understanding of the impacts of education policies and practices across countries, hence provide important clues to educational policy makers, administrators, teachers, and researchers to improve education quality. As TIMSS gains importance all around the world, the number of TIMSS studies is increasing in Turkey. Therefore, it is necessary to answer the question of which type of research was conducted related TIMSS so far. The purpose of this study is to reveal the tendency of all research done by TIMSS Turkish sample and classify them by using a developed framework. EBSCHO and ULAKBIM (Turkish National Database) databases scanned with the keyword "TIMSS" and all articles and thesis using Turkish sample were included in this study. 50 articles and thesis were found and classified according to the framework including three main categories: (1) Achievement: Whether the study examines science or mathematics achievement or both to explain factors affecting achievement. (2) Comparative Studies: Whether the study compares Turkey with other countries, or whether the study makes year comparisons, or both. (3) Item Evaluation: Whether the study evaluates items in terms of psychometric properties, or in terms of bias, or in terms of content and curriculum, or in terms of only descriptive properties, or in terms of item types. In these classifications, there were articles and thesis located in more than one category. The findings of this research can be illuminating as it enables to have a general overview about the research done in TIMSS so far, and it also give opportunity to see what further research can be done in this topic.

Identifying Top Talents within a Group of Successful Managers

Ole I. Iversen

BI Norwegian Business School, Oslo, Norway

e-mail: oii@assessit.no

Over the last decades we have become increasingly aware of the importance of leadership for organisational success (Hogan and Kaizer, 2005). A number of researchers focusing on personality traits have found the Five Factor Model (FFM) to be a useful model in predicting leadership effectiveness (Salgado 1997). This study has a closer look on a group of successful managers to se whether the FFM can be used to predict top talents within this group. The sample consists of 188 managers below 40 years identified by their manager as superior performers. A Norwegian version of NEO FFI (Martinsen et al. 2005) was used to measure personality. Data of job performance was collected by a ten items questionnaire (Kuvaas and Dysvik, 2009). Performance data were collected from three sources, self report, direct reports and superior. As

seen in Table 1, the managers had an average t-score on neuroticism about one SD below the norm group, and a score on extraversion and conscientiousness both about one SD above the norm group, and a score on openness and agreeableness which is more or less around the mean for the norm group, confirming results found in previous research (Salgado 1997). A correlation analysis between the performance measures and the FFM resulted in only three significant correlations out of 30 possible combinations (Table 2 and 3). The results indicates that the FFM might not be suitable to use to distinguish top talents from a group of successful managers, indicating that personality tests are better used in the first phase of the selection process. The findings also suggest that the higher score not necessarily is the better. Having a t-score between 50 and 70 on extraversion and conscientiousness and a t-score between 35 and 45 on neuroticism seems ideal.

Factor Structure of the Czech version of the Intelligence and Development Scales

Michal Jabůrek[1], Jan Širůček[1] and Tomáš Urbánek[2]

[1]Masaryk university, Faculty of Social Studies, Brno, Czech Republic; [2]Masaryk university, Faculty of Arts, Brno, Czech Republic

e-mail: michal.jaburek@gmail.com

The Intelligence and Development Scales - IDS (Grob, Hagmann-von Arx, Meyer, 2009) is a complex intelligence test for children (5 to 10 years). IDS consists of 19 subtests (21 in the Czech version) and focuses on 6 functional areas - Cognition, Psycho-Motor Skills, Social-Emotional Competences, Mathematics, Language, and Achievement Motivation. This study is a part of a broader standardization study of IDS, which involved 1455 children. The goals of this study are 1) to compare the factor structure of the original and Czech version, and 2) to verify meaningfulness of dividing the verbal and nonverbal subtests into 2 factors, which is based on the tradition of testing by Wechsler tests. To achieve both confirmatory factor analysis (maximum likelihood) was used. The four-factor structure (Extended Cognition, Psycho-motor Skills, Social-emotional Competence and Achievement Motivation) was confirmed. This structure, except for minor differences, corresponds with the original version. Two-factor solution (Verbal and Non-verbal Cognitive Abilities) was proved to be acceptable, but high correlation of these two latent variables ($r = 0.88$) has been found. Factor validity of the Czech version of the IDS was proved. Moreover, our results bear important consequences for consulting practice and interpretation of intelligence test results. High correlation between latent variables Verbal and Nonverbal cognitive abilities suggests that in the general population the verbal and nonverbal subtests of IDS measures two sides of the same disposition rather than two different abilities.

Validity Analysis of the Value-Added Indicators for Polish Schools

Aleksandra Jasinska and Anna Hawrot

Educational Research Institute, Warszawa, Poland

e-mail: a.jasinska@ibe.edu.pl

Assessment of school effectiveness requires good measurement tools. One such a tool - value-added indicators (or educational value-added - EVA) – is being developed in Poland. To ensure reliable and valid school effectiveness assessments EVA undergoes ongoing improvement and extensive quality check based on the following areas: verification of properties of achievement tests, adequacy of the statistical model, and validity of EVA estimates. The aim of the study was to assess the validity of EVA indicators for Polish lower secondary schools (LSS). We assumed that school effectiveness correlates with numerous processes and phenomena related to

students' social, emotional and cognitive development. Thus, we expected EVA estimates to correlate with pupils' relative intellectual growth during lower secondary schooling. We used data collected during two waves of the study on the development of EVA models conducted by Educational Research Institute (Poland). We analyzed data of over 5200 students from 291 classes and 150 schools gathered when pupils were in grade 1 and 3 of LSS (two intelligence measurements, controlled variables). Additionally, we acquired the results of national external examinations carried out just before the first grade (at the end of the primary school) and at the end of the third grade. It allowed us to estimate students' absolute and relative intelligence gains as well as to compute EVA indicators for sampled schools and to establish the status of the relationship between EVA indicators and students' intellectual development. The analyses indicated the existence of a correlation between EVA indicators and relative mean intelligence gains among students. The results support validity of Polish EVA indicators and prove utility of the chosen study design in checking their quality. They also give insight into patterns of human intellectual development.

Estimation of Reliability Coeficient for Ordinal Scale of Measurement: Scale on Study Skills and Strategies

Eva Jiménez García[1], Coral González Barbera[1], Eva Expósito Casa[2] and Esther López Martín[2]

[1]University Complutense of Madrid, Faculty of Education, Madrid, Spain; [2]National University of Distance Education. Faculty of Education, Madrid, Spain

e-mail: evajimenezgarcia@gmail.com

The most important psychometric properties of measurement instruments are the reliability and validity. There are many different ways to come up to their estimation, from considering conceptual and epistemological aspects until questions related properly speaking with the Classical Test Theory, from where addresses. Nevertheless, it is very common estimate the reliability coefficient based on internal consistency of these using the coefficient alpha de Cronbach, regardless of the measurement level of the variables and items who compose the scale. The aim of this study is to compare three methods for estimating the Reliability coefficient, with the goal of show that, in spite of the Alpha of Crombach is used almost indiscriminately, without regard to aspects like the measurement level of the variables, there are other more appropriate ways for the estimation of the reliability coefficients in the case of scales composed of ordinal items, as the Alpha Ordinal Coefficient and the Theta Ordinal. Both of them use the correlation polichoric matrix in the estimation (Sitjtsma, 2009). In order to achieve this objective, we estimate the three reliability coefficients pointed out previously in a adaptation of the Scale on Study Skills and Strategies- LASSI (Likert scale). The results show that the use of ordinal coefficients for the estimation of reliability with this type of scales, while taking into account the correct measurement level, offer higher values than those obtained with the traditional alpha (alpha Cronbach: 0,88; alpha ordinal: 0,93; theta ordinal: 0,92). The main conclusion of this study reinforces the idea of taking into account different levels of measurement for variables or items that are part of the measurement instruments when we select the procedure to estimate the reliability coefficient, inasmuch as it is not only more correct from the psychometric point of view but also we achieve that the coefficient rises.

Incidence of Behavioral and Emotional Problems among Brazilian Adolescents

Ana Paula Justo[1], Claudiane Aparecida Guimarães[2], Vivian Mascella[3], Mônica Maria Marques Suzigan[4] and Sônia Regina Fiorim Enumo[5]

[1]Pontifícia Universidade Católica de Campinas - São Paulo - Brasil, Nova Odessa, Brazil; [2]Pontifícia Universidade Católica de Campinas - São Paulo - Brasil, Uberlândia MG, Brazil; [3]Pontifícia Universidade Católica de Campinas - Sâo Paulo - Brasil, Sorocaba, Brazil; [4]Clinical Psychologist, Americana, Brazil; [5]Pontifícia Universidade Católica de Campinas - São Paulo - Brasil, Campinas, Brazil

e-mail: paulajusto@yahoo.com.br

Adolescence is a development transition period vulnerable to the occurrence of emotional/behavioral problems, likely due to the exposure of young people to a great number of stressors. In Brazil, there are few instruments to evaluate this population. Recently, the Youth Self-Report (YSR) was validated, still with a few studies. Using this instrument, this work appraised emotional/behavioral problems in a sample with 85 students (57 girls) from a public school at São Paulo, with 12-15 years old (M = 14), from 8º to 9º grade. The outcomes were classified by the Behavioral and Emotional Rating Scale (BERS), corresponding to the sum of all items; Internalizing Scale (IS), grouping items like anxiety/depression, withdrawal/depression and somatic complaints; and, Externalizing Scale (ES), grouping items like rule violation and aggressive behavior. The results are classified in 3 groups: clinical (enough emotional/behavioral problems so that there is clinical concern), borderline (a more detailed assessment is needed), and normal. For the research, the participants in the borderline group are arranged within the clinical group. Considering the Behavioral and Emotional Rating Scale (BERS), most adolescents (68%) were in the normal group (40 girls; 18 boys). However, 49% of the participants (27 girls; 15 boys) were at the clinical group for internalizing problems; and 25% of the participants (15 girls; 6 boys) were at the clinical group for externalizing problems. The frequency of emotional/behavioral problems at this sample was considered high, specially regarding internalizing problems (IP = 49%), mostly for not being an ambulatory sample. In the three scales of the YRS, girls are most likely to show problems. These outcomes reinforce the hypothesis of increasing adolescent vulnerability to the emerging emotional/behavioral problems and, taking into account the compromises those problems may cause to the development, they point out the need of preventive actions, early diagnosis, and appropriate treatment.

Validation of a Scoring Model for Short Adaptive Personality Questionnaires

Richard Justenhoven

cut-e GmbH, Hamburg, Germany

e-mail: richard.justenhoven@cut-e.com

The adalloc method (adaptive allocation of consent) is a scoring model for adaptive measurement of psychological constructs developed by Preuss (2002) with the aim of yielding differentiated profiles while keeping the questionnaire short. Adalloc strives to combine the advantages of normative and ipsative scoring models and overcome some of their limitations (for an overview see Murphy, Jako, & Anhalt, 1993; Zickar & Gibby, 2006; Bartram, 2006). The first sector of an instrument consists of a random sequence of one item per construct presented in randomly generated triplets. From the second sector onwards, triplets are combined on the basis of previous ratings so that it is not necessary to present all possible combinations of triplets, which allows for shortening the questionnaire (Preuss, 2002). The aim of this study is to identify

whether or not the random combination of items into triplets in the first sector impacts a candidate's profile because it is the basis for combining the triplets from the second sector onwards. This would impact construct validity. For testing the mentioned effects, data gathered using the adalloc-based questionnaire shapes (cut-e, 2002) of N=15,000 participants were analysed. The hypothesis was that if there was an impact of the initial combination of triplets, there would be significant differences between an item's mean scores, depending on which other items it would be combined with. The hypothesis was tested using ANOVAs. There were no significant differences between item mean scores across different combinations of items, meaning that the combination of triplets in the first sector does not impact the profile the questionnaire yields. The study can be seen as an indicator for validity of the method. However, further studies are required, particularly model-data fit analyses.

A Study of Analysis to the Performance of Preschoolers on Computerized Visual Perception Tests

Chin Kai Lin[1], Huey-Min Wu[2], Bor-Chen Kuo[3] and Yu-Mao Yang[4]

[1]National Taichung University of Education, Taichung, Taiwan; [2]Research Center for Testing and Assessment, National Academy for Educational Research, New Taipei City, Taiwan; [3]Graduate Institute of Educational Measurement and Statistic, National Taichung University of Education, Taichung, Taiwan; [4]Department of Language and Literacy, National Taichung University of Education, Taichung, Taiwan

e-mail: linchinkai97@gmail.com

The purpose of this study is to develop a "Computerized Visual Motor Integration Assessment Tool" for children from aged 4 to aged 6. The development of visual motor skills is required prior to children developing their writing skills, and the skill can help predict children's writing performance. Thus, the process of visual motor integration assessment can be used to assist in understanding children's fine-motor movement skills. Currently, all the popular Visual Motor assessment tools used in Taiwan are developed by overseas scholars; however, Chinese characters and English alphabets are greatly different. This study developed a set of computerized visual perception tests to evaluate children's ability to write Chinese characters. The designing process includes the consideration of the features of Chinese characters, thus dividing Chinese characters, based on their structures, into single characters and combination characters. The subjects are totaling 298 children, of which, 158 are male, and 140 are female. There are 30 items in this test, where 10 items are related to single characters, while 20 items are related to combination characters. The reliability index of this test is 0.83 (>0.7). Using the item response theory model to further explore the merits of the test questions, the results show that overall model fitting is good, and difficulty index are from 0.94 to 2.04.Lastly, One-way ANOVA is adopted to analyze whether age and gender have any impacts on writing performance. The analysis results show that the effectiveness of the test is better for children between the ages of 5 and 6 than children of 4 years old, and that there is a significant difference in the results. However, there is no significant difference in writing performance in terms of gender. The computerized visual perception integration assessment tool can effectively determine the writing performance of children.

Re-Test-Reliability and Stability of the Hand Preference Test HAPT 4-6

Ursula Kastner-Koller[1], Pia Deimann[1] and Johanna Bruckner-Feld[2]
[1]Univeristy of Vienna, Vienna, Austria; [2]University of Rostock, Rostock, Germany
e-mail: ursula.kastner-koller@univie.ac.at

The Hand Preference Test for 4- to 6-year-olds (HAPT 4-6; Bruckner, Deimann, & Kastner-Koller, 2011) is a new observational method to assess handedness and hand preference. Compared to existing tests for handedness, its added value is an explicit and detailed assessment of hand preference and consistency of hand preference at a very young age. So far, internal consistency and validity of the test scores (Laterality Quotient LQ and Consistency of Hand Preference CHP) were analyzed, both yielded high coefficients. With regard to the development of handedness and hand preference in early childhood, analyzing the re-test reliability and the stability is of particular interest. Participants were 44 children (21 boys, 23 girls) who had to accomplish the HAPT 4-6 thrice (at the age of four, five and six years). In addition the QHP-Task (Quantification of Hand Preference Task, Bishop et al., 1996), a hand preference test, and the PTK-LTD (Schilling, 2009), a performance test, were applied. Hand preference (LQ) proved to be stable whereas consistency of hand preference (CHP) varied over time. Moderate to high correlation coefficients between HAPT 4-6, QHP-Task and PTK-LTD furnish further evidence for the concurrent validity of the HAPT 4-6.

Measurement of Emotional and Behavioral Disorders in Adolescents by Using Latent Classification Models

Sung Eun Kim[1] and Seulki Koo[2]
Ewha Womans University, Seoul, South Korea
e-mail: miilli@hanmail.net

The present study aims to make a diagnosis of emotional and behavioral disorders in adolescents based on latent classification models. In the present study performed in 3 steps as research process. In the first step, we develop the Q-matrix. The First draft of Q-matrix is constituted through factor analysis and then, the 5 experts refine the draft. After the experts' discussion(or specialists' conferences), statistical Q-matrix validation(Chiu, 2013) is applied to revise the Q-matrix. The 5 experts discussed the revised Q-matrix to decide the final Q-matrix. As the seconds step, it is analysis by G-DINA(generalized deterministic input; noisy "and" gate; de la Torre, 2011) model. Then, in order to assessing the model fit, we compute the goodness-of-fit indices. In the last step, we interpret the results and provide an example in which we propose a new form of reporting the results to present a diagnosis of psychological problems. In this study, the response data of The Korean Children and Youth Pannel Survey(KYCPS) were used. KCYPS selected 21 questions constructed 3 factors and were collected response targeting 2,351 people the eighth grade. The questions are 'attention deficit and hyperactivity'(7), 'aggressive behavior'(6), 'somatoform'(8) of the three emotions. In the results of the study, we can investigate the behavioral disorders that have the highest retention rate of youth in Korea, and the overall trend that is most collective potential student belongs to what groups by classifying the potential groups. In addition, the diagnostic results of individual student\'s emotional and behavioral status will provide a tailored prescription. This researched we performed was used for the cognitive diagnostic model mainly, however, we hope that it will become more widely used in psychological field in the future as well.

Psychometric Properties of Travelers Personal Questionnaire

Jelena Kolesnikova[1], Jelena Levina[2] and Tatjana Turilova-Miscenko[3]
[1]Riga Stradins University, Riga, Latvia; [2]International Higher School of Practical Psychology, Russia;
[3]University of Latvia, Riga, Latvia
e-mail: jkolesnikova@inbox.lv

The purpose of this research was to develop the Russian version of Travelers Personal Questionnaire(TPQ) that measures travelers' personal characteristics and to determine its psychometric properties. The TNQ was developed for travelers with native Russian language from different countries: Russia, the Baltic states, Belarus, Ukraine.The sample consisted of 160 participants aged from 18 to 68 years (M = 31.17, SD = 9.29) (male – 44%, female – 56%). The factorial validity of the TPQ was established using principal components analysis with varimax rotation; this yielded five factors: Social Impact (k = 6), Dependence (k = 6), Sensation Seeking (k = 7), Conservatism (k = 6), Sociability (k = 5). All the TPQ scales had high internal consistency (Cronbach's alpha variedfrom .79 to .84).The reaction and discrimination indices satisfied the accepted psychometriccriteria. The psychometric properties of TNQ satisfied criteria. The further stage of TNQ development would be the confirmatory factor analysis in broader international sample, the concurrent and convergent validity establishing, and test-retest reliability examination.

Differential Item Functioning across Test Versions

Maciej Koniewski, Przemyslaw Majkut and Paulina Skórska
Educational Research Institute, Warszawa, Poland
e-mail: maciej.koniewski@uj.edu.pl

In Poland national exams are administered to students finishing primary school (13 y.o.), lower secondary (16 y.o.), and higher secondary school (19-21 y.o.). The first high stake exam is being administered after lower secondary school. In these pen and paper exams, the test is delivered in two test forms to reduce the risk of cheating. Both test forms share the same items, but the sequence of options for the multiple choice items differ across forms. The study aims to assess differential item functioning (DIF) across the test forms. During the item and response analyses we observed unexpected differences in students performance across two test forms. The biggest differences were observed for items bundled in testlet, where the correct answer pattern for items within a testlet are marked with the same key (e.g., A, A, A) in one version and in the alternative booklet the key pattern is mixed (e.g., A, B, C). The two test forms could be viewed as theoretically parallel in terms of test items, but that difference in key patterns might influence students' performance on the items. DIF analyses across different forms of test booklets were performed on 2013 data from version A and B of 'The history and knowledge about society' test in lower secondary schools in lubelskie, malopolskie and podkarpackie voivodeships (n = 81 545). To detect DIF, the Mantel-Haenshel test, logistic regression and standardization were used. Methods of DIF visualization were demonstrated. Results indicated meaningful differences in items functioning across test forms, especially when the pattern of keys on items bundled in testlet were marked with the same symbol. Such a pattern has been named an 'anti-pattern', because test takers might consider such response pattern very unlikely and in consequence provide wrong answers. Findings of this study would provide important guidelines for test development.

On Designing Data-Sampling for Rasch Model Calibrating an Achievement Test

Klaus D. Kubinger[1], Dieter Rasch[2] and Takuya Yanagida[3]

[1]University of Vienna, Faculty of Psychology, Division of Psychological Assessment and Applied Psychometrics, Maria Anzbach, Austria; [2]Institute of Applied Statistics and Computing, University of Natural Resources and Applied Life Sciences, Rostock, Germany; [3]University of Vienna, Faculty of Psychology, Division of Psychological Assessment and Applied Psychometrics, Salzburg, Austria

e-mail: klaus.kubinger@univie.ac.at

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. In analogy to Andersen's approach in his Likelihood-Ratio-test, the testees must be divided into at least two groups according to some criterion suspected of causing differential item functioning (DIF). Then a three-way analysis of variance design with mixed classification is the result: There is a (fixed) group factor A, a (random) factor B of testees within A, and a (fixed) factor C of items cross-classified with ; obviously the factor B is nested within A. Yet the data are dichotomous (a testee either solves an item or fails to solve it) and only one observation per cell exists. The latter is not assumed to do harm, though the design is a mixed classification. But the former suggests the need to perform a simulation study in order to test whether the type-I-risk holds. 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.

Somatic-Affective and Cognitive Depressive Symptoms among Older Finnish Natives and Somali Refugees Measured by the Beck Depression Inventory

Saija Kuittinen

University of Tampere, Tampere, Finland

e-mail: saija-liisa.kuittinen@helsinki.fi

Epidemiological studies reveal that depression is globally the most common psychiatric disorder (Bromet et al., 2011). However, the frequency of specific depressive symptoms, such as bodily pains or excessive self-criticism, varies according to ethnic groups (Weissman et al., 1996). Therefore, Western based scales may not be sensitive to culturally salient forms of depression across different populations. Our study adds to the research by first, comparing the manifestation of depressive symptoms between two underrepresented groups in the literature: older Somali refugees and native Finns, and second, by using the original version of the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). Our research questions are the following: Can the BDI scale be used to identify different types of symptom manifestation? And if so, do the Somalis and Finns differ in their symptom manifestation? The participants are 256 match-paired older Somali refugees and Finnish natives derived from national register. The depressive symptoms were first conceptually classified into Cognitive and Somatic-affective categories. A Principal Axis Factor analysis partly confirmed the conceptual categorization, although there were some BDI items that were problematic in the two factor solution. There were 10 final items in the Somatic-affective symptoms category, e.g., Fatigue.

The Cognitive symptoms category includes eight items that represent self-centered negative thoughts, e.g., Pessimism. The proportionality coefficients (Tucker's phi) for the symptom categories were both above .85 (Ten Berge, 1986) and for the Cognitive symptoms above the commonly accepted value of .90 (He & van de Vivjer, 2012), thus allowing further comparisons. Results showed that the Somalis manifested more Somatic-affective symptoms of depression than the Finns, whereas the Finns manifested more Cognitive symptoms than the Somalis. The results indicate that awareness of the variation in symptom manifestation is crucial when assessing depression in cross-cultural contexts.

An Examination of the Short-Term Stability of a Measure of Inspection Time

Joseph Kush

Duquesne University, Pittsburgh, United States

e-mail: kush@duq.edu

Mental chronometry, is considered to be an index of the speed and efficiency of cognitive processing and numerous research studies have found a relationship between the speed and efficiency of cognitive processing and intelligence. Measures of Inspection-Time (IT) have been found to correlate with psychometric general intelligence yet minimize, the role of prior knowledge or achievement-related content. These tasks are seen as distinct from those found on traditional IQ tests, however they are thought to assess the same features of cognition that underlie the performance of more complex intelligence test tasks. There currently exists very little psychometric data to support IT methodologies. While some Cronbach alpha reliability estimates have been calculated, very little reliability information in IT measures can be found. Therefore the goal of this study was to provide this type of psychometric support; specifically test-retest reliability. One hundred thirty five university students served as participants. Students were tested individually, on two occasions, following a two-week interval. Participants completed a PC-based, single response, Inspection Time task requiring a two-alternative, forced choice response. When first and second administrations of the IT task were compared, the differences between response times was not significantly different ($p = .367$). Test-retest correlations, based on a two-week interval, were adequate for the traditional pi stimulus ($r = .72$, $p < .01$). The current study offers some preliminary psychometric evidence in support of a computer-based IT measure. In interpreting the current findings, it was surprising that the coefficient was not higher given that the test-retest interval was only two weeks. Additional research must be conducted, in this regard, across more diverse populations including varying age-rages, levels of cognitive ability, and test-retest intervals.

Personality Profiles and Social Skills in Adolescents

Ana Betina Lacunza[1], Evangelina Norma Contini De González[2] and Claudia Paola Coronel[3]

[1]Consejo Nacional de Investigaciones Cientificas y Técnicas CONICET, San Miguel de Tucuman, Argentina; [2]Facultad de Psicología. Universidad Nacional de Tucumán, San Miguel de Tucumán, Argentina; [3]Facultad de Psicología. Universidad Nacional de Tucumán, Yerba Buena-Tucumán, Argentina

e-mail: betinalacu@hotmail.com

There is controversy over the impact of personality traits on social behaviour. It is argued that personality by itself is not enough to guarantee an adequate social adjustment and that social skills are presented as an adaptation factor to it. The of this work were a) to describe the social skills in adolescents who attend school in San Miguel de Tucumán, Argentina, and b) to establish relations between personality dimensions and social skills in the participants. A quantitative

study with a comparative and transversal design and a non-probability and purposive sampling was carried out. 260 adolescents between 12 and 15 years old who attend state and private schools were administered the Eysenck Personality Questionnaire (EPQ-J), the Silvia and Martorell Socialization Battery (BAS-3) and a sociodemographic survey. The results showed that deficits in social skills were related to withdrawal and social anxiety in adolescents from state schools, while their peers from private schools showed deficits in leadership. Regarding personality traits, adolescents with lower consideration of others showed lower extraversion and higher emotional toughness (psychoticism), a trait which is also associated with adolescents who lack social self-control. Emotional instability (neuroticism) was related to withdrawal and social anxiety, while low perception of leadership was associated with lower extraversion. The data described emphasize the bidirectional relations between social behaviour and personality during a stage of identity consolidation in adolescents, in particular. Moreover, it is highlighted the importance of delving into the role of context in shaping personality profiles with psychopathological tendencies related to social deficits.

Gender and Mathematic Competence Achievement

Ainhitze Larrañaga and Paula Elosua
Euskal Herriko Unibertsitatea EHU/UPV, Donostia, Spain
e-mail: ainhilarra@gmail.com

Gender difference in children's math competence is been deeply investigated in the last decades. Some studies suggest that differences between boys and girls start at the first years of school, however some researches claim that the gender difference appears in High School, and even many studies report that the difference in mathematical competence is not significant. In this controversial contexts the goal of this research was to analyze gender differences in a math competence test. The test is composed by 29 items which cover four different math dimensions: (1) quantity, (2) space and shape, (3) changes , relationships and uncertainty and (4) problem solving. The sample consists of 18,045 Elementary School students - girls 8,771 (48.6%) and 9,274 boys (51.4 %) - aged between 10 and 11 years. We analyzed the psychometric properties of the test according to the gender, and compared the achievement by dimension. The results showed slight differences in mathematical competence in the dimensions of (1) quantity, (2) space and form, and (4) problem solving, although the effect size associated with them was not high.

Math Test Achievement According to Linguistic Variables: Family Language and School Language

Ainhitze Larrañaga and Paula Elosua
Euskal Herriko Unibertsitatea EHU/UPV, Donostia, Spain
e-mail: ainhilarra@gmail.com

In bilingual educational contexts text language can influence in the academic achievement. The skills acquired in the first language (L1) and second language (L2) influence directly student's cognitive development and thus, in the academic performance. Traditionally, it has been suggested that the acquisition of math competence literacy stayed on the sidelines of the vehicular language. However, nowadays, it is suggested that vehicular language´s influence affects directly. In this sense, the purpose of this study is to analyze the math performance taking account of of linguistic variables: familiar language and school language. The data came from a math competence test composed by 29 items. The sample consists of 18,045 students of 11 years old enrolled in the Basque Country. Analyses were performed in

the context of the general linear model. The results show the relationship between the analyzed linguistic variables and the performance in math.

Evaluating Chinese-Language Versions of Numeracy Scales: A Study from Taiwan

Joseph Lavallee[1] and Supin Hung[2]

[1]Ming Chuan University, Taipei, Taiwan; [2]National Cheng Kung University, Tainan, Taiwan

e-mail: lavallee@mail.mcu.edu.tw

Low numeracy skills have been associated with suboptimal decision-making in domains such as personal health and finance. Recent years have seen a growth of interest in this topic and the appearance of a number of numeracy scales. In addition to language issues, a difficulty in using these scales in E. Asian contexts is that population numeracy skills there tend to be higher, threatening the ability of such scales to discriminate across samples. Two scales, the abbreviated Rasch numeracy scale (Weller et al., 2012) and the Berlin Numeracy Test (Cokely et al., 2012), have been proposed for use in discriminating across broader ability ranges. The purpose of the present study was to evaluate the performance of translated versions of these scales with a Taiwanese sample. Twenty-two items from six (partially overlapping) numeracy scales were first translated into Chinese (traditional script) through an iterative back-translation procedure. The resulting instrument was then administered to 221 university students in Taiwan. Best results were obtained by combining the 4-item Berlin Numeracy Test with the 3-item General Numeracy Scale (Schwartz, 2003): basic psychometric indicators (mean = 3.71, SD = 1.72, Cronbach's alpha = .64, mean inter-item correlation = .35) were consistent with reported results from the original samples, and the distribution of scores was the closest to normal (skewness = -.237) of the instruments compared in the study. The results offer support for the use of Chinese-language versions of numeracy scales originally developed in English for western populations. The discussion turns to further steps underway to validate these scales with a broader sample in Taiwan. The problem of language-related differential item functioning will also be raised in the context of the current study.

Psychometric Properties of Short Versions of the Travelers Needs and Personal Characteristics Questionnaires

Jelena Levina[1], Jelena Kolesnikova[2] and Tatjana Turilova-Miscenko[3]

[1]International Higher School of Practical Psychology, Russia; [2]Riga Stradinš University, Riga, Latvia; [3]University of Latvia, Riga, Latvia

e-mail: elena.levina@inbox.lv

The purpose of this research was to introduce the short versions of the Travelers Needs Questionnaire and the Travelers Personal Characteristics Questionnaire. Based on the Travelers Needs Questionnaire (TNQ) and the Travelers Personal Characteristics Questionnaire (TPQ) archive, 25 of 53 items were selected to construct a new short form of the Travelers Needs Questionnaire (TNQ-S) and 20 of 30 items were selected to construct a new short form of the Travelers Personal Characteristics Questionnaire (TPQ-S). The sample consisted of 160 participants aged from 18 to 68 years (M = 31.17, SD = 9.29) (male – 44%, female – 56%). The factorial validity of the TNQ-S was established using principal components analysis with varimax rotation; this yielded seven factors: Personal Development (k = 4,), Physical Hedonism (k = 4), Pilgrimage (k = 4), Cultural Development (k = 3), Professional Realization (k = 3), Communication and Social Recognition (k = 4), Sport (k = 3). KMO was. 79. All the TNQ-S scales had high internal consistency (Cronbach's alpha varied from .73 to .90). The factorial validity of the TPQ-S was established also using principal components analysis with varimax rotation; this

yielded five factors: Social Impact (k = 4), Dependence (k = 4), Sensation Seeking (k = 4), Conservatism (k = 4), Sociability (k = 4). KMO was. 79. All the TPQ-S scales had high internal consistency (Cronbach's alpha varied from .80 to .84). The reaction and discrimination indices of TNQ-S and TPQ-S satisfied the accepted psychometric criteria. The psychometric properties of TNQ-S and TPQ-S satisfied the criteria. The further stage of TNQ-S and TPQ-S development would be the confirmatory factor analysis in broader international sample, the concurrent and convergent validity establishing, and test-retest reliability examination.

Multiple Standard Setting Study Outcomes: Empirical Exercises Informing Theory

Gad Lim[1] and Chad Buckendahl[2]

[1]Cambridge English Language Assessment, Cambridge, United Kingdom; [2]Alpine Testing Solutions, Orem, UT, United States

e-mail: lim.g@cambridgeenglish.org

Obtaining different outcomes on different occasions has been observed in standard setting (SS). This may not be considered a problem because SS is currently conceptualized as a policy-driven exercise. However, many SS situations are concerned less with policy but with some set of criteria or standards, i.e. not what level of the standard is acceptable but rather what the standard is in the first place. This makes it more of an objective data-related enterprise, and which SS theory does not adequately account for at present. The objective of the paper is to illuminate SS theory to account for situations which are objective data-constrained enterprises, arguing that convergent outcomes should be expected. The paper will offer a procedure and guidelines for determining recommended cut scores based on triangulation and accounting for measurement error. Two separate SS exercises were conducted relating a large scale, international test of English language proficiency, to the Canadian Language Benchmarks, which describes different levels of language ability, using appropriate methodology for different subtests (Yes-No Angoff, Analytical Judgment). Results from the two exercises were then compared. Results show a high level of agreement in the outcomes of the two studies, especially when taking SEM into account. In cases of disagreement, a consistent adjudication procedure suggested itself for application to produce realistic cut scores. In SS situations not driven primarily by policy, the task is to find the cut score objectively corresponding to the standard involved. Multiple SS exercises should be seen as repeated measures whose combined results provide a more reliable cut point estimate. Convergent outcomes should be expected, and divergent outcomes investigated to determine their sources. This approach brings SS closer in line with other areas of assessment, subject to validation and falsification, leading to better-supported validity arguments.

The Classification Validity of DINA Model in Complicated Cognitive Diagnostic Assessment

Zhe Lin, Wei Tian and Tao Xin

Beijing Normal University, Beijing, China

e-mail: lz_psy@163.com

The cognitive diagnostic assessment based on cognitive diagnostic modeling?CDM? has received extensive attentions for decades, for its utility in identifying whether individuals has mastered specific knowledge or skills. DINA model is a simplified but interpretable CDM that has been widely applied to real diagnostic assessments. However some problems still remained. One of the problem is that how well DINA model could accurately classify the individuals' knowledge state (KS) . Classification validity is undoubtedly the prerequisite for the explanation of the KS and instructing remedial teaching. However, most researches could only give an estimated value to evaluate the classification validity due to the unavailability of true knowledge state.

Furthermore, It is still unknown whether such simple DINA model could Make accurate classifications in a more complicated diagnostic assessment which contains a complex Q-matrix with a hierarchy of attributes. This study examined the DINA's classification validity and influential factors through a Primary School Arithmetic Word Problem.1240 students participated in the test. All items were dichotomously scored. The Q-matrix and attribute hierarchy were constructed based on experts' discussion and students' think-aloud data. For some items examined about 4-8 attributes, so the test is complicate. The most outstanding advantage of this research was that students'attribute mastery patterns on each item were given by their teachers, which can be used as true KS. Using CDM package in R program to calibrate parameters and estimate individuals' KS. And according to estimated and true KS, some statistics like RMSE, MAD and IMstats could be calculated. all of them are various evidences reflecting the classification validity. results showed that 1.RMSE associated with each attribute fell within 0.1-0.3; 2. the extremely easy or difficult items may have a misleading influence on the classification accuracy; 3. classification accuracy would be higher when attribute hierarchy was taken into consideration.

Psychometric Properties of the Formal Characteristics of Behaviour-Temperament

Wen Liu

Liaoning Normal University, Dalian, China

e-mail: wenliu703@126.com

This paper presents a Chinese adaption of the Formal Characteristics of the Behaviour-Temperament Inventory (FCB-TI), a self-report instrument that evaluates six temperamental scales, based on Strelau's concept of temperament. A first sample of 626 undergraduates completed the Chinese version of the Regulative Theory of Temperament Questionnaire (RTTQ), which is an initial pool of 381 items. Internal consistency suggests adequate reliability, and an exploratory factor analysis (EFA) revealed a 6-factor solution consistent with the original instrument. The CFA revealed good support for the temperament structure with a second sample of students (N=2980). Internal consistency and factorial structure were re-examined and test-retest correlations over a two-week period were calculated with a third sample of adults (N=2265). Convergent and discriminant validity was explored in relation to EPQ model dimensions. Findings indicate that the Chinese version of the FCB-TI has similar psychometric properties and generally satisfactory reliability and validity.

The Effects of Personality and Network Structure on Local Network Accuracy

Jing Liu, Min-Qiang Zhang and Wen-Qing Tang

Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou, China

e-mail: jing.liu1219@gmail.com

Local network accuracy refers to an accurate perception of the other person's perception of his relationship with you. People differ in their ability to perceive accurately the group's network structure. Existing research leaves us largely unable to explain the variation in accuracy in social network perception. To explore the claim that local social network perception should consider the effect of both situational factors and individuals' personality differences, data of 632 students from 20 classes in 4 different universities were collected by using EPQ-RSC, Rosenberg self-esteem scale and cognitive social structure questionnaire. The study adopted cognitive social structure (CSS) and multilevel structural equation model (MSEM) to investigate the effect that personality (extraversion, psychoticism and self-esteem) and network structure (centrality,

cognitive centrality and network density) may have on local network accuracy. The result showed that individual who had higher level of self-esteem, centrality and network density tended to have more accurate local network cognitive, while psychoticism and cognitive centrality had the opposite influence. It was also evidenced that centrality and cognitive centrality were not related. But that extraversion related to local network accuracy was not supported. This study indicated that 1) low self-esteem individuals' lack of stability of their self-concept could cause low local network accuracy, 2) high centrality individuals had more source of information which facilitated them to know about their relationships better, 3) denser networks provide its member with more information about its local structure and tended to be more stable which cost them less effort to know about their local structure, 4) high cognitive centrality could give people more confidence in their popularity and reduce their attention to relationship and thus decrease the accuracy of perception, 5) perception of local network may be more influenced by mechanism of cognitive process rather than local structure itself.

Improving Test Performance by Cognitive Training: Implications for Employment Testing

Katharina Lochner and Achim Preuss

cut-e Group, Hamburg, Germany

e-mail: katharina.lochner@cut-e.com

For many years, it has been widely acknowledged that performance on tests assessing general mental abilities (GMA) is relatively stable and can only marginally be improved by training. However, recent research questions this notion. Firstly, performance on tests assessing GMA such as Raven's APM improved with repeated testing (e.g. Bors & Vigneau, 2003) or with training the working memory (e.g. Jaeggi, Buschkuehl, Jonides, & Perrig, 2008). Secondly, in the assessment of Learning Potential, a test-train-retest approach is used on purpose (e.g. de Beer, 2005). Both approaches have in common the finding that performance on cognitive tests can be improved significantly. The purpose of the present study was to investigate whether training would improve performance on computer adaptive tests. We analysed data from a cognitive training program with more than 50,000 participants of different ages, educational backgrounds and occupations. Participants repeatedly completed tests assessing reasoning, short term memory, and concentration in unsupervised mode. All tests in the programme use item generators. Individuals' performance improved a standard deviation on all three tests after 20 repetitions. Training caused to a stable significant upward trend in test performance. Limitations of the study are that it was conducted in unsupervised mode and that an analysis of transfer effects was not possible. However, it has some potential implications for employment testing that will be discussed in the session. What do the results mean for the reliability and validity of tests? Will a trained person also perform better on the job than an untrained person? How can we assess a person's training level, and how can we take different training levels into account in a selection setting?

The Impact of Specific Positive and Negative Emotions on the Performance on an IQ test

Katharina Lochner[1], Michael Eid[2] and Achim Preuss[1]

[1]cut-e Group, Hamburg, Germany; [2]Freie Universität Berlin, Berlin, Germany

e-mail: katharina.lochner@cut-e.com

Tests assessing general mental abilities are good predictors of future job performance (Schmidt & Hunter, 1998). The assumption is that the performance on such a test is determined by the test taker's cognitive ability. However, studies found mood to influence performance on analytical tasks (e.g. Abele, 1995; Knapp, 1988). The purpose of the study was to systematically examine

the impact of specific emotions (joy, sadness, anger, composure) on analytical tasks, as proposed e.g. by Lyubomirsky, King, & Diener (2005). In an unsupervised online experiment, N=429 participants of various ages, educational backgrounds, and occupations first completed a test assessing reasoning ability. Then they were randomly assigned to one of five conditions and watched a film clip inducing joy, composure, sadness, anger, or a neutral state (Gross & Levenson, 1995). Afterwards, they completed a parallel version of the reasoning test. Data were analysed using two-factorial mixed ANOVA and structural equation modelling. There was a main effect of the within person factor with a significant improvement from the first to the second test, but there was no main effect of the group factor (emotional state), nor was there an interaction effect of within and between person factor. In the present study, emotional state did not impact test performance. It is the first study that investigated specific positive and negative emotions in an experimental setting. Furthermore, participants were rather diverse. However, as it was conducted in unsupervised mode, a replication under supervised conditions would be desirable. Moreover, participation in the experiment was voluntary, therefore self-selection effects are likely. Finally, the question is whether the results can be generalised to other tests assessing the same construct.

A Bootstrap Method to Replication: The Bootstrap Replicability Coefficient p_rep^B

Man Lok Chan

The Chinese University Of Hong Kong, Hong Kong

e-mail: mlchan@psy.cuhk.edu.hk

The probability of replication prep, introduced by Peter Killeen, is the probability of replicating an effect of an experiment. It aims to be an alternative to null-hypothesis significance tests. Nevertheless, Since it has been established and promoted in statistics inference in psychology studies, its validity has been questioned by several researchers. Regarding the bootstrap resampling method, it is a procedure by drawing successive samples with replacement from an observed data set repeatedly to obtain a virtual population. In this study, based on the Killeen's concept of the probability of replication, the bootstrap procedure is proposed to compute the Bootstrap Replication Index p_rep^B. A simulation study was conducted to examine the empirical performance of p_rep^B under different levels of population effect sizes and sample sizes. Result indicated that given the same effect size and sample sizes, p_rep^B is more accurate than prep to predict the probability of replicating an effect of an experiment. To conclude, p_rep^B outperforms Killeen's prep.

TESIS and MSCEIT: Convergence (or Divergence) between Two Different Ability-Based Approaches for Assessing Emotional Intelligence

José Héctor Lozano Bleda[1], Jorge Barraca Mairal[1], Antonio Fernández González[2] and Héctor Opazo Carvajal[2]

[1]Universidad Camilo José Cela, Madrid, Spain; [2]Universidad Autónoma de Madrid, Madrid, Spain

e-mail: jhlozano@ucjc.edu

The assessment of Emotional Intelligence (EI) through ability-based tests allows for an objective measurement free from the biases associated with self-report. The aim of this study was to explore the convergence between two different EI ability tests: the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCETI – Mayer, Salovey, & Caruso, 2001) and the Sensitivity to Social Interaction Test (TESIS – Barraca, Fernández-González, & Sueiro, 2009). The MSCEIT, on the one hand, aims to measure the four branches of the Mayer-Salovey model (i.e., Emotional Perception, Emotional Facilitation, Emotional Understanding, and Emotional Management). The TESIS, on the other, is aimed to assess the ability to discriminate emotions in social contexts (the Emotional Perception and Emotional Understanding branches of the Mayer-Salovey model).

Whereas the tasks of the MSCEIT are mostly based on static visual information, the items in the TESIS consist of cinematographic scenes based on audiovisual information in real time. A correlational study was conducted in which both instruments were applied to a sample of 164 undergraduates from 18 to 37 years old. Contrary to expectations, the TESIS showed significant correlations with the Emotional Facilitation and Emotional Management indices of the MSCEIT, but not with the Emotional Perception and Emotional Understanding indices. Both tests also diverged in their estimate of individual differences according to age, sex, and academic background. A more in-depth analysis of the results revealed that the significant correlations between both tests corresponded to those tasks in the MSCEIT that involve discriminating the function played by emotions in defined context. On the contrary, correlations were not significant with those MSCEIT tasks consisting just in identifying emotions from decontextualized information. Our results lead us to encourage the development of multiple EI ability-tests based on contextualized audiovisual information in order to analyze their convergence and ultimately better understand the nature of the EI construct.

The Trees: Simple Visual Discrimination Test (DiViSA): Evidence of convergent validity with measures of impulsivity and attention from the Faces', Differences Perception Test and the d2 Test of Attention

José Héctor Lozano Bleda, Elena Capote Calvo and María Poveda Fernández Martín

Universidad Camilo José Cela, Madrid, Spain

e-mail: jhlozano@ucjc.edu

The Trees: Simple Visual Discrimination Test (DiViSA – Santacreu, Shih, & Quiroga, 2010) is a computerized test developed to assess the inattention and impulsivity components contemplated in the DSM-V for the diagnosis of AD/HD in children. As a visual discrimination test, the DiViSA has the advantage of being free from the biases and low interjudge reliability associated with rating-scales, and also shows advantages over Continuous Performance Tests derived from the simultaneous presentation of stimuli. The DiViSA has already shown appropriate psychometric properties in terms of both reliability and validity (Quiroga, Santacreu, Montoro, Martínez-Molina, & Shih, 2011; Santacreu et al., 2010). The aim of this study was to obtain additional evidence of convergent validity for the measures of the DiViSA in relation to other well-established visual discrimination tests for assessing attention and inhibitory control in children: the 'Faces', Differences Perception Test (Thurstone & Yela, 1995) and the d2 Test of Attention (Brickenkamp, 2002). To that end, a correlational study was carried out in which the aforementioned instruments were administered to a sample of 111 students from primary education (46 males and 65 females) between the ages of 8 and 12 years. The correlation analysis showed values in line with expectations among the three tests. The results of the MANOVA revealed that the tests also converged in their estimate of the individual differences in attention and impulsivity according to the age and sex of the participants. The results provide new evidence of convergent validity for the DiViSA in relation to the Faces and the d2, in comparison to which the DiViSA would offer certain advantages due to its computerized format.

A Forced-Choice Test for Assessing Work-Related Competencies

Beatriz Lucia[1], Vicente Ponsoda[2], Francisco José Abad[2], Daniel Morillo[2], Iwin Leenen[3], Alejandro Vidal[2] and Sonia Rodriguez[1]

[1]Instituto de Ingeniería del Conocimiento, Madrid, Spain; [2]Universidad Autónoma de Madrid, Madrid, Spain; [3]Universidad Nacional Autónoma de México, México D.F., Mexico

e-mail: beatriz.lucia@iic.uam.es

Forced-choice items allow some control of response biases, such as social desirability, but the data they provide may be ipsative and have certain peculiarities that hinder its psychometric analysis. To overcome these limitations two IRT-based models have been proposed: the Multi-Unidimensional Pairwise-Preference Model (MUPP; Stark, Chernyshenko, & Drasgow, 2005), and the Thurstonian Item Response Model (TIRT; Brown and Maydeu-Olivares, 2011). The aim of the study was the development of a forced-choice test to measure four work-related competencies (work planning and scheduling, focusing on efficiency, problem solving, and management of change) based on the TIRT model. Four initial scales, one for each of the previous competencies, have been developed and applied to a sample of about 2000 candidates/incumbents, in a normative format. IRT assumptions were checked and items calibrated under the two parameter logistic model. A simulation is being conducted to select the forced-choice format (blocks of either 2 or 3 items) and, from each initial scale, the 12 items that better fulfill the requirements of an as-efficient-as-possible forced-choice test of 24 or 16 blocks (for the 2- and 3-item blocks, respectively). The forced-choice test will be administered to a new sample of candidates/incumbents, and the results submitted to calibration and psychometric analysis. The whole set of 48 items will be administered to the same new sample as well in a dichotomous normative format. The reliabilities of the scale scores and the person parameters obtained under the single-stimulus and forced-choice formats will be compared. Special attention will be paid to the two correlation-between-competencies matrices, as any difference may indicate the advantage of the forced-choice format in the control of the social desirability bias.

Automatic Item Generation for Number Series Reasoning Test

Fang Luo[1], Yanan Liu[1], Yunyun Zhang[2] and Hongyun Liu[1]

[1]School of Psychology, Beijing Normal University, Beijing, China; [2]Faculty of Education, Beijing Normal University, Beijing, China

e-mail: luof@bnu.edu.cn

In order to accomplish the automatic item generation of the number series, the stimulus features of the items should be defined to explore their effect on the difficulty of number series items. In this study, operation categories and operand categories were selected as generating components, and periodicity, working memory capacity, and operand size were quality control components. Based on the combination of the generating components, two kinds of number series were generated, which were named single operation number series items and double operation number series items. 469 students had the testing of these items, some of that were from secondary vocational school and the others of that were from some college. The results showed that single operation number series items were so easy that they were only suitable for the low ability students; the double operation number series items were appropriate for measuring the inductive ability of college students. Operation categories and operand categories as well as their interaction affected the difficulty of the single operation number series items. Operation categories of each layer affected the difficulty of the items, and their interaction effect was not found. Furthermore, some cognitively diagnostic models were used to estimate the

difficulty of any generating component. The difficulty of new item automatically generated could also be predicted with these models.

Dimensionality Issues Related to Item Format in Large Scale Achievement Tests

Davide Marengo[1], Michele Settanni[1] and Renato Miceli[2]

[1]University of Turin, Turin, Italy; [2]University of Turin, University of Aosta Valley, Turin, Italy

e-mail: davide.marengo@unito.it

Numerous studies suggest that while multiple-choice (MC) and constructed-response (CR) items for the most part measure the same construct, a residual local dependence among CR items producing unintended multidimensionality can be observed in mixed-format tests. Still, mixed-format tests are commonly intended to provide a single measure of ability. Using data from a standardized mixed-format math test, we tested the hypothesis that a two-dimensional compensatory IRT model letting MC and CR items load on a common primary math dimension, and the CR items additionally load on an a secondary dimension would be more appropriate than a unidimensional model. As a second aim, the distinctiveness of the two dimensions was investigated by controlling for ability differences according the relevant grouping variables (gender, citizenship status and regularity in studies). As a final aim, we investigated the degree agreement between the two models in the classification of students in different proficiency levels. Analyses were performed on random sample data (N=3000) coming from the 8th grade high-stake math examination implemented in Italy as part of the INVALSI 2012 state-wide assessment program. The existence of a secondary CR ability dimension was not fully supported by the results. While an increase in fit was found implementing the two-dimensional model, the low variability observed on the secondary CR dimension suggested the violation of unidimensionality to be negligible. No significant differences across the two dimensions were found comparing students' ability on the grouping variables. Finally, a strong concordance was found between the tested models in the classification of students in proficiency levels. Our results appear to be in line with findings reporting overall construct equivalence between the abilities assessed by CR and MC items when CR items are mainly designed as short-answer items as opposed to open-ended questions.

Confirmatory Factor Analysis of the Cattell-Horn-Carroll-Based Dutch Cognitive Ability Test

Marlies Tierens [1], Magez Walter[2], Annemie Bos[1] and Decaluwé Veerle[1]

[1]Thomas More University College - Applied Psychology, Antwerp, Belgium; [2]Coordination Team Antwerp for Psychodiagnostics (CAP vzw), Brasschaat, Belgium

e-mail: marlies.tierens@thomasmore.be

Recent cognitive instruments are often, explicitly or implicitly, based on the cognitive abilities from the Cattell – Horn – Carroll (CHC) taxonomy. The Dutch Cognitive Ability Test (CoVaT -CHC) is a new CHC-based intelligence battery for children and adolescents in Flanders (Dutch speaking part of Belgium). The purpose of the test is to provide insight in general intelligence as well as specific individual cognitive strengths and weaknesses. It measures five broad cognitive abilities (Fluid Intelligence, Crystallized Intelligence, Short-term Memory, Visual Processing and Processing Speed) and consists of both verbal and language-reduced subtests thereby making the test useful for non-native Dutch speakers. The test can be used for individual or group assessment. The purpose of this study is to evaluate psychometric properties of the CoVaT-CHC. Confirmatory factor analysis will be conducted to determine which model best describes the structure measured by the CoVaT-CHC: the hierarchical CHC-model, the Gf-Gc model or a model similar to Spearman's g. The representative sample consisted of approximately 2000 children

who all completed the CoVaT-CHC. Participants ranged from 10 years, 0 months to 13 years, 11 months. Participating children completed the assessment at school in groups of 5 – 25 during four sequential lessons. Results & will be presented at the conference and will be discussed in the context of previous findings. Preliminary analyses support the structure of the CHC-model and the subtests show good to excellent internal consistency.

Neuropsychological Assessment in the Elderly: a Systematic Review of Brazilian Studies

Vivian Mascella[1], Ana Paula Justo[2], Claudiane Aparecida Guimaraes[3], Luiz Ricardo Vieira Gonzaga[4], Vanessa Marques Gibran[5] and Sonia Regina Fiorim Enumo[4]

[1]Pontifícia Universidade Católica de Campinas - São Paulo, Sorocaba, Brazil; [2]Pontifícia Universidade Católica de Campinas - São Paulo - Brasil, Nova Odessa, Brazil; [3]Pontifícia Universidade Católica de Campinas - São Paulo, Uberlândia, Brazil; [4]Pontifícia Universidade Católica de Campinas - São Paulo, Campinas, Brazil; [5]Pontifícia Universidade Católica de Campinas - São Paulo, Americana, Brazil

e-mail: vivian.mascella@hotmail.com

Neuropsychological assessment is a way of research cognitive and behavior functions, in increasing use, indication and studies. It has been used in different samples, especially assessment in the elderly. This research achieved a systematic review about empiric studies of neuropsychological assessment in the elderly, regarding the year of the review, body of knowledge, content, and most used neuropsychological tests. The database from the Scientific Eletronic Library on Line (SciELO-Br) was consulted, using "assessment" and "neuropsychological" as keywords. It was found 127 published reviews between June 1987 and May 2013; 110 were discarded by study type, sample's age group appraised, and nonspecific tests used in the study. Therefore, 17 empirical studies with a sample above the age of 65 years old, published between 2001 and 2013, especially 2008, were read and analyzed. The reviews are mainly in the field of Medicine (70.6%) and Psychology (29.4%). It was found a range of topics (15), more often verifying psychometric property testing in the elderly, dementia, cognitive performance, education level influence, executive function, cognitive abilities in elderly widowers and epidemiological studies. It was identified 43 instruments, including: Mini-Mental State Examination (MMSE), which is an instrument for cognitive screening; Trail Making Test, which evaluates cognitive attention and flexibility; Rey Auditory-Verbal Learning Test (RAVLT), which appraises learning and memory; and, Clock Drawing Test, which is also an instrument of cognitive screening. As a conclusion, Psychology could function and publish more about Neuropsychological Assessment, and invest more in the creation and validation of neuropsychological tests for this sample in the country.

Croatian Standardization of the Minnesota Multiphasic Personality Inventory-2 - Restructured Form (MMPI-2-RF)

Krunoslav Matešic[1], Krunoslav Matešic, Jr.[2] and Valentina Ruzic[3]

[1]Faculty of Humanity and Social Sciences, Zagreb, Croatia; [2]Croatian Catholic University, Zagreb, Croatia; [3]Naklada Slap, Zagreb, Croatia

e-mail: kmatesic@nakladaslap.com

MMPI-2-RF is the revised, shortened version of the MMPI-2, published in 2008 (University of Minnesota Press). It is composed of 338 true-false items selected from the MMPI-2 to reflect contemporary theories and models of psychopathology and personality. The MMPI-2-RF inventory is designed to assess clinically relevant variables and the advantage of this form of inventory is the smaller number of items and shorter administration time, compared to the MMPI-2 inventory. It is intended for testing persons over the age of 18. Administration usually

takes 35 to 50 minutes. This revised form of inventory includes empirically valid scales which provide information about clinical symptoms, personality characteristics, behavior, interpersonal functioning and interests of the person being assessed. The objective of the study was to develop the Croatian edition of MMPI-2- RF inventory which will contain Croatian norms. The MMPI-2-RF normative sample is drawn from the MMPI-2 normative sample and consists of 587 subjects (327 F and 260 M), aged 18 to 86, from all parts of Croatia. The inventory was administered individually and answers were combined in 51 scales: 9 validity scales, 3 Higher-Order Scales, 9 Restructured Clinical Scales, 23 Specific Problems Scales, 2 Interest Scales and 5 Personality Psychopathology Scales. Norms were calculated using linear and uniform T-scores, the same as in U.S. standardization. The result of the standardization is the Croatian edition of the MMPI-2-RF inventory published by Naklada Slap in 2014. It is expected that this version of the MMPI inventory will be very useful in a variety of situations: in clinics, in diagnosing mental disorders and in selecting appropriate treatment for every patient, helping clinicians in Croatia diagnose and treat patients.

Classroom Culture and Educational Climate Within and Between Countries: Links to Student Mathematics Achievement in PISA

Carina McCormick[1], Sara E. Gonzalez[1], Leslie R. Hawley[2] and Betty Jean Usher-Tate[1]

[1]University of Nebraska-Lincoln, Lincoln, NE, United States; [2]The Nebraska Center for Research on Children, Youth, Families & Schools (CYFS), Lincoln, NE, United States

e-mail: mccormick@huskers.unl.edu

Classroom culture created by the teacher is linked to achievement in mathematics (see Hiebert & Grouws, 2007). Policy makers often turn to results from Programme for International Student Assessment (PISA) for guidance about reforms as lower-performing countries attempt to emulate educational practices of higher-performing countries. In interpreting such results there is often a confounding of effects at distinctive hierarchical levels: the country level, the classroom level, and the student level (see Raudenbush & Bryk, 2002). When research results shape policy and practice, it is essential that generalizations about the relationship between classroom culture and student achievement are appropriately interpreted at the correct level. The current study examines the links between classroom culture and teacher practices with student achievement in mathematics using multilevel modeling. The study separates the variance of the predictors into within-school, between-school, and between-country components with the outcome of 2012 PISA mathematics achievement, incorporating plausible values methodology. Using a three-level model with the HPMIXED procedure in SAS 9.3, predictors for the models are index variables focusing on teacher practices, student-teacher relationships, discipline, engagement, and motivation, as well as student income. Preliminary results that compared a model of country means to a model of schools within the United States showed vast differences in the pattern of results, supporting the need to more explicitly address the patterns of relationships across and within countries for the full PISA data set. Much of the needed preparation for such models have already been completed in the preliminary analysis stages. Full results will be presented at the conference. The results of this study will be of immediate policy relevance as discussions regarding optimal classroom practices, particularly in mathematics and science, increasingly gain prominence. In addition, this study illustrates the value of PISA data in addressing complex educational issues.

Validity Evidences for an Electronic Scale of Attitudes Towards Statistics via Item Response Theory

Claudette Maria Medeiros Vendramini[1] and Camila Cardoso Camilo[2]

[1]Universidade São Francisco, Itatiba/SP, Brazil; [2]Universidade São Francisco, Amparo/SP, Brazil

e-mail: cvendramini@uol.com.br

The aim of this study was to analyse the validity evidences based on the internal structure of the electronic scale of Attitudes toward Statistics – eEAEst through Rasch partial credit model, one of the favorite models for the application in the field of attitudes measurement. This scale measures the students' attitudes toward statistics, defined as a disposition to respond favorably or unfavorably to situations related to statistics learning. A sample of 183 students, selected for convenience, participated in this study. They were aged 18-58, M=28.6 and SD=7.9, 60.1% female, enrolled in different semesters from different graduation courses of a private university in the interior of São Paulo state, Brazil, and had already attended a Statistics subject. The participants answered the 36-item, 5-point Likert-type scale, in computer laboratories, after having agreed on a Free and Clarified Consent Term. The results indicated both a good fit of the items to Rasch partial credit model and the scale's good psychometric properties, for a scale composed with making it a good instrument to measure the construct attitude. The sample's average attitude towards statistics was more positive rather than negative. This study can contribute to the improvement of teaching and learning of statistics and help other areas that need statistical concepts.

Psychometric Properties of a Brazilian National Exam of Pedagogy Students' Performance via Item Response Theory

Claudette Maria Medeiros Vendramini and Fernanda Luzia Lopes

Universidade São Francisco, Itatiba/SP, Brazil

e-mail: cvendramini@uol.com.br

A large scale assessment, in the educational context, aims at investigating the students' abilities and competences through their several education stages. One of these tests supposed to assess the specific knowledge within distinct knowledge areas is the National Student Performance Exam (ENADE), taken by Brazilian students since 2004. The aim of this study is to assess the psychometric properties of the ENADE's Pedagogy exam which took place in 2011, by means of Rasch Model of the Item Response Theory. We used a data base containing academic records of 86,844 Brazilian Pedagogy students who took the exam in 2011. They were aged 19-80, M=33.5 and SD=9.2, 93.3% female. The students answered both the general formation and the specific component tests containing multiple choice and essay questions. The results suggested items with good psychometric properties, and the item difficulty coefficients were, in average, of intermediate difficulty. Most students showed an ability level compatible to the average difficulty of the items, indicating that the exam has a difficulty level which is similar to the students' ability level. This study may contribute to develop studies that investigate the quality of higher education assessment, since this will lead to constructing education based on better properties and having a better social, political, and economical recognition.

Self-Efficay Beliefs, Strain, and Personality among University Students

Enrique Merino-Tejedor
University Of Valladolid, Segovia, Spain
e-mail: enmerino@psi.uva.es

Well-being is an issue of growing interest for researchers in psychology and other fields of social sciences. It is important to identify those hurdles that can hinder to attain an adequate state of well-being; strain can become an important obstacle when university students are thriving to get their academic and career goals. In this study we tried to deepen into the state of strain among university students and its relationship to personality variables and self-efficacy beliefs. A sample of 206 university students from different grades participated in this study. Self-efficacy beliefs were assessed through the University Self-Efficacy Scale; several instruments were used in order to assess strain: The Scholar Burnout Inventory (SBI), the Irritation Scale for university settings, and other instruments to appraise the state of fatigue. Finally, personality dimensions were assessed through the Overall Personality Assessment Scale (OPERAS), an instrument inspired in the Five Factor Model (FFM) of personality, which assess the following five dimensions: Extroversion, agreeableness, conscientiousness, neuroticism, and openness to experience. Results obtained showed interesting and significant results in the way expected. For example, we found significant correlations between some personality variables (extroversion, neuroticism, and agreeableness) and academic irritation. Furthermore, significant and negative correlations were obtained between all the five dimensions of personality and the state of fatigue, assessed through the Fatigue Assessment Scale (FAS). On the other side, results yielded significant and negative correlations between self-efficacy beliefs and burnout, irritation and fatigue, as predicted by the theoretical model. Besides, self-efficacy beliefs correlated in a positive way with the five personality dimensions of the FFM. Results found in this study suggest important lines of investigation for future research, as for example to design programs in order to improve self-efficacy beliefs among university students to lessen their level of strain and burnout.

Adaptation of the WIAT-III Oral Subscales on a Cypriot Sample

Michalis Michaelides, Andry Vrachimi-Souroulla, Chrysanthi Leonidou and Georgia Panayiotou
Dept. of Psychology, University of Cyprus, Nicosia, Cyprus;
e-mail: michalim@ucy.ac.cy

The Wechsler Individual Achievement Test–Third Edition (WIAT-III) is an individually administered assessment designed to measure students' reading, listening, speaking, writing and mathematics skills. In the latest version, the 16-subscale test was revised to provide comprehensive achievement information for the purposes of identifying student strengths and weaknesses, determining eligibility for academic services, and planning and implementing teaching interventions (Breaux, 2009). This research examines the adaptation in Greek of the five Oral Language subscales: Receptive Vocabulary, Oral Discourse Comprehension, Expressive Vocabulary, Oral Word Fluency, and Sentence Repetition. The subscales were adapted in Greek and administered to a sample of students from Cyprus (N=539) in grades pre-K to 12. A clinical subsample (N=43) included participants diagnosed with mental retardation, ADHD, learning or developmental disabilities. All five subscales had high, statistically significant inter-correlations ranging from .52 to .74. Cronbach's alphas ranged between .67 and .90. Females scored higher on average than males on all subscales, although the mean differences were non-significant. Comparisons between the non-clinical and the clinical samples revealed significant differences on four of the subscales with medium effect sizes (.25<d<="" p=""></d.

Everybody Lies

Iva Mikulic[1], Ana Simunic[2], Ana Prorokovic[2] and Ljiljana Gregov[2]
[1]Sunce International Health Centre, Zagreb, Croatia; [2]University of Zadar, Zadar, Croatia
e-mail: iva.mikulic@live.com

The general purpose of this study was to bring contributions to the findings about the validity of using lie scales in selection situations as a measurement or indicator, to which psychologists base there decisions concerning one's abilities. Because selective situations are subject to (di)simulation, it is assumed that lie scales on personality questionnaires do not contribute to objective decision concerning psychological abilities of a candidate. A sample of 666 candidates was identified during selection procedures. The aim was to examine whether candidates who have been diagnosed with mental disorders, which is a contraindication for driving and other activities, lie more than candidates with no diagnosis by using the EPQ (Eysenck, H.J., Eysenck, S. B.G., 1975). As was expected, candidates with no mental disorders lie equally as the candidates with diagnosed mental disorders do. In both groups (Men, N=548) significant correlations were found between Lie and Neuroticism scale. Some studies show that a significant negative correlation between the neuroticism and lie scale indicate a high motivation for (di)simulation (Levin and Montag, 1987; Cowles et al., 1992; Jackson and Francis, 1998) and that in such circumstances, the lie scale can be used as a tool to seclude (di)simulators, so as to abstract 5% of the highest results on the lie scale. However, in order to determine an appropriate threshold for exclusion of individuals it is necessary to have data on the total population as well as the participants' age (Eysenck and Eysenck, 1994). Even though it is assumed that people who have been diagnosed have higher motivation for (di)simulation because of the objective reason, in selective situations 'everybody lies'. It is questionable if it is justified using such scales as any kind of measurement psychologists use to decide if a certain candidate is suitable to drive, to work, etc.

A New approach for Modeling Local Item Dependencies in the C-test

Dorothea Mildner[1], Johannes Hartig[1] and Andreas Gold[2]
[1]German Institute for International Educational Research, Frankfurt, Germany; [2]Goethe University, Frankfurt, Germany
e-mail: mildner@dipf.de

The conditional independence assumption in educational measurement requires test items to be independent conditional on the latent trait. Items sharing a common passage as in testlets may violate this assumption. The C-test, where items are embedded in the passage itself, is a testlet-based integrative testing instrument that measures overall language competence. Two forms of local item dependence potentially occur in the C-Test: a) trait dependence, where subsets of items have varying levels of dependence between their underlying traits, and b) response dependence, where knowing the answer to one item increases the probability of knowing the answer to the next item. A new approach for modeling these local item dependencies in the C-test was developed to investigate if a distinction between the two forms of item dependencies is actually possible, and if so, both forms are empirically distinguishable when occurring simultaneously. Data from a C-Test that measured general competence in English as a foreign language (N = 9816 ninth graders in Germany) was used to answer these questions. Four different psychometric models were compared: a unidimensional model, a testlet model (modeling trait dependence), a response dependence model (modeling response dependence),

and a hybrid model (modeling both trait and response dependence). These item response models have not yet been compared to each other; especially the response dependence model, as well as the hybrid model, has not yet been estimated directly. Results showed that the hybrid model fitted the data best, followed by the response dependence model and the testlet model. The unidimensional model showed the poorest fit. Thus, we can conclude that local item dependencies exist in the C-test, and that the two forms are distinguishable even if they occur simultaneously – like in the C-test. Implications for the analysis of testlets in general and C-tests in particular are discussed.

Developing Computer Version of Schwartz's Portrait Value Questionnaire Revised (PVQR) Translated in Russian

Olga Mitina, Veronika Sorokina and Lena Rasskazova
Lomonosov MSU, Moscow State University of Psychology and Education, Moscow, Russia
e-mail: omitina@inbox.ru

The PVQR is well known questionnaire translated in many languages (including Russian) and used in many countries including comparative large scale surveys. It allows determining the importance for a person of 19 values. Computer technology allows to make testing more easy and more effective. The computer version was created, which differed in the form of presentation of questions (the screen), and in the way of fixing the answers. Respondents expressed extent of their agreement or disagreement using a slider and moved it among a continuous scale. The program recorded the position of the slider along the interval from 0 to 100.Access to the survey was realized through the internet, testing can be conducted in any place. Data are stored on a remote server. The current study was conducted during September - November 2013, it was attended about 700 people (388 females, 286 males) aged 15 - 17 years, studying in public schools. Psychometric characteristics of the items and the scales of PVQR were compared with similar characteristics obtained from the results of a standard blank survey, which was conducted at the same time among similar respondents. Values which got the largest preferences were benevolence, independence, hedonism and reputation. Least preferences were got by conformism and power. Comparisons between boys and girls showed that there is no difference in the hierarchy of values, but there are statistically significant differences in absolute scores. Comparing with boys girls showed relatively greater desire for benevolence, tolerance, boys relatively higher desired to dominate. The research battery also included test for determining self-control, behavioral styles (J.Kuhl), competence in managing (Nizovskikh, Mitina, Tiulkin), time perspective inventory (Zimbardo), life satisfaction scale (Diener). These scales were used for correlational validation of the test. Analysis of correlational and mean structures in female and male samples separately was done.

Indigenous Student's Educational Conditions Associated to the Admission Tests of the Higher Education in Costa Rica: A Multi-Level Analysis

Tania Elena Moreira Mora
Instituto Tecnológico de Costa Rica, Cartago, Costa Rica
e-mail: tmoreira02@hotmail.com

A group of researchers at the Instituto Tecnológico de Costa Rica (TEC) and the Universidad de Costa Rica (UCR) posed the question if the indigenous student's cultural, linguistic and educational features could affect their performance on the admission tests for these state universities. The aim of the study was to estimate the degree and nature of some of the main educational disadvantages of the indigenous groups in their access to the higher education. This

is a co-relational exploratory and descriptive research administered to a sample of 82 students in four different indigenous high schools, plus one rural (n=18) and other urban (n=171). The data collection was held from 2011 to 2012. In the multilevel regression model, the first level variables were their age, traveling time, native language, self-classification ethnic, kind of school, and ethnic identity. The second level was the group's cluster. The dependent variables were the scores gotten from two specific tests and the admission tests (PAA) of the UCR and TEC. In the PAA of TEC, the explained variance (R-squared) was 23.6% and the predictors with statistical significance were kind of school (Coef.-6.17) and the student's age(coef.-.699). In the PAA-UCR no variable turned out to be statistically significant. In the Language test, the R-squared was 0.1233 and age was the only evidence (Coef. -1.43) and in the reasoning test with figures (PRF), the explained variance was 29.45%, the high school´s kind (coef. -4.927015) and the students' age (coef. -1.71) were associated to a level of alpha=0.05. The students from indigenous high schools, who are above the average age, tend to get a lower score at the admission tests. That is why their probabilities to access to a public college are less than the rural and urban group.

A Best-Practice Overview of Techniques to Analyse Holland's Circumplex Vocational Personality Model

Brandon Morgan and Gideon De Bruin

University of Johannesburg, Johannesburg, South Africa

e-mail: bm.morgan@yahoo.com

Holland's influential circular/hexagonal vocational personality theory specifies six types that form a circumplex (Realistic, Investigative, Artistic, Social, Enterprising, Conventional). Conventional factor analytic techniques should not be applied to the investigation of their structure. Rather, a variety of circular/circumplex techniques that are either descriptive or confirmatory should be used: multidimensional scaling (MDS), the randomisation test of hypothesised order relations (RTHOR) and covariance structural modelling (CSM). This presentation provides a critical overview of these techniques and demonstrates their application to measuring Holland's circumplex model. The aims of this study were to provide a critical overview of techniques used to analyse circumplex models; and to demonstrate the application of these techniques to Holland's circumplex model. A total of 985 participants completed the South African Career Interest Inventory. The participants were obtained from three higher education institutions across various fields of study. MDS is a descriptive non-parametric technique used to visually investigate circular ordering. While it was readily used, recent practice is to use confirmatory approaches that provide an indication of model fit, such as the RTHOR and CSM. The RTHOR typically produces better fitting models than CSM because it is non-parametric. However, CSM provides indices of model fit and can measure models with different constraints. While each of the techniques has limitations, best practice incorporates all of them to obtain different perspectives on the data. In the data MDS demonstrated a circular ordering of the six types. The RTHOR demonstrated good fit while CSM indicated acceptable to poor fit depending on the constraints that were specified. These results show that relying on only one technique can produce misleading inferences. Various circumplex analysis techniques should be applied to investigate Holland's circumplex model rather than relying on conventional factor analytic techniques. The data demonstrated the strength and weaknesses of the different approaches.

Structural Equation Modeling Based Reliability Estimations

Josu Mujika and Paula Elosua
Euskal Herriko Unibertsitatea EHU/UPV, Donostia, Spain
e-mail: josu.mugica@ehu.es

Defined as the proportion of test score variance associated with true score variance in Classical Test Theory, test score reliability is one of the most reported statistics in social sciences. The first choice of the applied researchers to estimate the reliability of total scores is by assessing covariability among items. More specifically, computing coefficient alpha. Nevertheless, its proper use requires compliance with strict conditions which in practice are rarely satisfied. As a result, in recent years many specialized researchers and journals have been interested in developing new approaches to reliability. This paper aims to describe one of those, Structural Equation Modeling (SEM) based estimation, and to compare to coefficient alpha. Estimation of reliability based on SEM allows to consider a variety of models and choose the one that best fits the data, including multi-dimensional structures, non-continuous responses and non-tau-equivalent items. However SEM based estimations require a model correctly specified and large sample size. In addition, the literature on the SEM based estimates of reliability is less extensive than the literature on alpha.

The Effect of Model Misspecification and Sample Size on Nonlinear SEM Estimates of Reliability

Josu Mujika and Paula Elosua
Euskal Herriko Unibertsitatea EHU/UPV, Donostia, Spain
e-mail: josu.mugica@ehu.es

During the last years Structural Equation Modeling (SEM) based estimates of reliability have captured the attention of specialized journals and researchers. Some studies carried out with continuous nature data warn about the sensitivity of these approaches when small sample sizes and incorrectly specified models are used. The present study aims to explore the effect of sample size and model misspecification on nonlinear SEM estimates of reliability. Analyzed data were generated by Monte Carlo simulation according to (a) four models: tau-equivalent items, tau-equivalent items with correlated errors, congeneric items and congeneric items with correlated errors; and (b) three sample sizes: 150, 300 and 1000. Nonlinear SEM estimates of reliability was obtained after fitting data samples to each one of the four models. The 48 experimental conditions (4 x 3 x 4) was replicated 1000 times. As expected, average bias was higher with small sample size and missespecified models than with big sample size and correctly specified models. In general terms, estimations introducing congeneric models produced less bias and confident intervals than tau-equivalent models.

Cognitive Ability Testing in Modern Societies

Claire Muller[1], Romain Martin[2], Franzis Preckel[3] and Tanja Gabriele Baudson[3]
[1]University of Luxembourg, Luxembourg, [2]University of Luxembourg, Walferdange, Luxembourg; [3]Universität Trier, Trier, Germany
e-mail: cl.muller@uni.lu

Thanks to recent technological advancements, the field of psychometrics currently faces the threshold to a new era of evaluation and assessment. The use of tablets with their intuitive tactile mode of operation, in combination with increased flexibility for the representation of task

elements (instruction, stimuli, answer mode, etc.), offers excellent new possibilities for testing at all ages. Potential advantages include easier, faster and cheaper test administration and evaluation, increased test taker motivation, increased sustainability, easy collection of process data, etc. Further, special relevance can be accredited to the possibility to create assessment procedures that could increase testing fairness for students with a minority background. Considering that modern societies are increasingly heterogeneous, this would certainly be an important feature. To date, however, empirical data in this domain remains scarce. Neither do we know in how far tablet-based assessment differs from traditional paper & pencil administration, nor have the effects of these different modes of test instruction and administration been examined for diverse populations. In our study, we therefore investigated the usefulness of these new technologies through systematic variation of administration modes for a series of cognitive ability tasks. Possible modes were tablet-based versus paper & pencil administration and 3 different methods of instruction: no instruction, verbal, and video-animation. In order to determine whether modern technologies prove fairer, we investigated whether test results changed depending on language background. Our sample consisted of 500 10-year-old students in year 4 (Cycle 3.2) of the Luxembourgish primary school, half of the students had a minority language background. Results of the different conditions of administration will be presented and discussed.

Assessment of Self-Regulation Factors for Training Transfer in Spanish Workers

Mariel Fernanda Musso[1], Carla Quesada[2], Anna Ciraso[2] and Eduardo C. Cascallar[3]

[1]UADE (Universidad Argentina de la Empresa)- K Universiteit Leuven, Buenos Aires, Argentina; [2]Universitat Autònoma de Barcelona, Barcelona, Spain; [3]K. Universiteit Leuven, Brussels, Belgium

e-mail: marimusso@uade.edu.ar

There is a large body of literature establishing the importance of motivational factors for learning and performance in different tasks. There is a need to validate the use of instruments in studies on self-regulated learning (SRL), examining their internal consistency and construct validity (Richardson, 2004), particularly when applied in different settings or socio-cultural groups. The objective of this study was to validate a self- report instrument to assess SRL, the On-line Motivation Questionnaire, part 1 (OMQ91; Boekaerts, 2002). This instrument was administered before the implementation of an employee training program in a workplace setting. A total of 2745 workers, of both genders (Female: 50.8%), from private (85.1%) and public companies, participated. This widely used self-report questionnaire consists of 23 items which measure three constructs: appraisals, emotions, and learning intention. The OMQ91 instrument was adapted and translated following the ITC Guidelines (ITC, 2010). An Exploratory Factor Analysis was conducted using a randomized sample (n= 846), applying a Maximum Likelihood extraction method with Promax rotation. The correlation matrix presented satisfactory values (KMO= .935; Sphericity test = 5537.441; p= .000). A structure of three factors explaining 49.35% of the variance was found. Factor 1 involved items related to "Learning Intention/Personal Relevance of the Task"; the second factor involved items regarding "Facility to apply", and the third factor consisted of items related to "Subjective Competence". Cronbach's alpha analyses showed good reliability for each scale. Finally, a Confirmatory Factor Analysis was carried out on a new randomized sample (n= 1882 participants). The three-factor model was found to have a good fit (897.176; df = 87; p= .000; NFI= .918; CFI= .925; RMSR= .07). These results will be discussed in the context of current SRL models, and of transfer of training studies in the work environment.

Prenatal Bond Assessment Scale (PBAS): Results of a pilot study

Lucía Navarro Aresti[1], Ana Martínez Pampliega[1], Ioseba Iraurgi Castillo[1] and Sagrario Martín Íñigo[2]
[1]University of Deusto, Bilbao, Spain; [2]Quirón Bilbao, Bilbao, Spain
e-mail: lucia.navarro@deusto.es

The study of prenatal bond -one of the main predictors of mother-child attachment created after birth- enables to detect possible difficulties for the mother when establishing an affective relationship with the fetus. However, it does not exist any suitable instrument that evaluates this union in a specific way and with good psychometric properties for the Spanish population. 525 pregnant women attending Maternal Education classes answered to the Prenatal Bond Assessment Scale (PBAS). This Likert scale has 56 items with five options for each answer divided in seven subscales: Search for information about the fetus, Fantasize about the fetus, Interact with the fetus, Fetal protection, Recognition of the fetus as a separate individual from the mother, Detection and gratification of the needs of the fetus, and Fear of loss of the fetus. All items have been reached by the judgment of three experts. The PBAS final version is composed of 24 items: 32 items have been removed due to their inadequate psychometric properties. Two dimensions, Fetal protection and Detection and gratification the needs of the fetus, have been grouped into one: Altruistic protection of the fetus. The global reliability of the inventory is high (0,83) and the dimension ratios that compose it oscillate between 0,60 and 0,86. Using Confirmatory Factor Analysis has been proved a model of six dimensions subsumed under a second order general factor, and has been obtained an adequacy fit indexes (□² = 502,94; p< .001; GFI= .92; CFI= .95; RMSEA= .03 [.02 - .04]). Due to its adequate psychometric properties, the PBAS can be proposed as a valid instrument to measure the prenatal bonding of the pregnant woman with the fetus.

Mother´s Perceived Behavioral Control of Child´s Fruit and Vegetables Consumption Questionnaire

Gabriela Navarro Contreras, Monica Fulgencio Juarez and Ferran Padros Blazquez
Facultad de Psicología. Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacán, Mexico
e-mail: gabriela.navarro.c@gmail.com

The rising of childhood obesity has increased the focus on the role of parents in the causes as well as its prevention. The theory of planned behavior considers the effect of perceived control on actual weight reduction, understanding that perceived control is likely to take into account some of the realistic constraints that may exist in intentions to lose weight (Schifter & Ajzen, 1985). According to Fishbein & Ajzen(2010) given a sufficient degree of actual control over the behavior, people are expected to carry out their intentions when the opportunity arises. Therefore, following the Theory of Planned Behavior (Ajzen, 1988, 1991) the objective of this study was to construct a questionnaire to asses Mother´s Perceived behavioral control of children ages 6 to 12 fruit and vegetables consumption. The first version of the questionnaire consisted in 9 forced choice items tapping parental beliefs about the presence of factors that may facilitate or impede performance of the behavior of child´s fruit and vegetables consumption. Items were revised and exploratory factor analysis and internal consistencies analysis were conducted using a sample of 140 mothers. Exploratory factor analyses tested a two factor model, which explains 52.53% of variance. Internal consistencies form the two factors were above .731. Further work with the instrument is needed to establish its reliability and validity. The questionnaire provides a tool for assessing Mother´s Perceived behavioral control of child´s fruit and vegetables consumption.

Multidimensional IRT Analysis of Large Scale Data in NCT

Tatsuo Otsu
Research division/NCUEE (alias DNC), Tokyo, Japan
e-mail: otsu@rd.dnc.ac.jp

The National Center Test (NCT) is a nationwide university admission examination conducted by NCUEE in January every year. More than a half million applicants take NCT every year. English is one of the most important subject of NCT. NCUEE introduced English listening comprehension test that use personal IC players in 2006. The Paper&Pencil section of English examination contains a subsection about English pronunciation. Items in this subsection ask for knowledge on phonetic symbols of English words. The purpose of this study is to show the values of these subsections for assessments. We analysed responses of English tests using multidimensional IRT models. The P&P section, including the pronunciation knowledge subsection, contained about 50 items. The pronunciation subsection were composed of seven items. The listening comprehension section contained 25 items. We assumed 2PL model with two latent factors. One factor corresponds to P&P items, and another corresponds to listening comprehension items. The items in the pronunciation subsection are admitted to have loadings from both factors. Computation with IRTpro2.1 on Xeon E5-1660 (six core CPU) took about seven hours for data of a half million test takers. The items in the pronunciation knowledge subsection show intermediate properties between other two item groups. The estimated correlations between the P&P factor and the listening comprehension factor were stable at 0.9 for years. The factor loadings for items in the pronunciation subsection, however, varied between years. Two dimensional IRT modes was useful for showing latent structures of large scale examination data. The most serious problem in practical use is the computational cost. We plan to show computation enhancement by GPGPU for fast estimation.

Hopeful People are Also More Satisfied?

Juliana Pacico[1], Micheline Bastianello[1], Ana Claudia Vazquez[2] and Claudio Hutz[1]
[1]Federal University of Rio Grande do Sul, Porto Alegre, Brazil; [2]Federal University of medical sciences of Porto Alegre, Porto Alegre, Brazil
e-mail: jucerentini@hotmail.com

Satisfaction of life and hope are strongly related to physical and mental health. Life satisfaction is a construct that refers to the cognitive evaluation that individuals do about their overall quality of life or on specific areas, such as work, leisure, love, health, and finance. Hope was defined as a positive motivational state that emerges from the interaction between pathways and agency when there is an objective to achieve. This goal should be important enough for people to feel motivated to overcome the obstacles that possibly will arise between them and their goal. Some studies indicate that life satisfaction is related to hope. The aim of this study was to investigate the correlation between hope and life satisfaction in Brazilian adults. The sample was composed by 524 Brazilian students, 17 to 36 years-old (M = 21, SD = 3.2), 57% women. The instruments used were the Adult Dispositional Hope Scale and the Life Satisfaction Scale. Both instruments were adapted and validated for use in Brazil. Results showed a significant correlation (r = .50) between hope and life satisfaction. This correlation is similar to those found in other studies (around r=.41) this result suggests that hopeful adults are more satisfied with their lives. A possible explanation for this finding would be that hopeful persons might draw multiple pathways to achieve their goals and that might make it easier for them to attain them. Therefore, they will be more successful, and success can produce greater satisfaction with life. These results also

suggest that if it is possible to develop interventions to make adults more hopeful these interventions may also increase life satisfaction.

Dispositional and Cognitive Hope through the Lifespan

Juliana Pacico[1], Ana Claudia Vazquez[2] and Claudio Hutz[1]

[1]Federal University of Rio Grande do Sul, Porto Alegre, Brazil; [2]Federal University of medical sciences of Porto Alegre, Porto Alegre, Brazil

e-mail: jucerentini@hotmail.com

There are different conceptions of hope. Dispositional hope is related to how individuals wish to achieve desired goals. It has a cognitive and an emotional component. Pathways, the cognitive component, are the ways that lead the person to an objective. Agency, the motivational component, uses pathways to drive the search for the objective. Cognitive hope is the interaction between wishes and expectations. It is related to future expectations, and has two subscales: hope-self (refers to hope about the subject itself), hope-other (refers to hope about others and global circumstances). The literatures refers that development of hope starts at the childhood and goes on through lifespan. The objective of this study is to present the relationship between dispositional hope and cognitive hope in a group of adolescents and to assess whether these results could be replicated in a group of adults. The participants of the first study were 368 adolescents (57.6 % female), mean age 15.9 years (SD=.81), and, of the second study, 849 undergraduates (57.4% female), mean age 21.2 (SD = 4.0). All subjects completed a sociodemographic questionnaire, the Adult Dispositional Hope Scale and The Hope Index. Pearson correlations between the two types of hope were performed. The results indicated that there is a higher correlation between dispositional hope and hope-self in the adolescent and in the adult (r=.42) groups. The results suggest that dispositional hope relates more strongly to hope- self because both have items that assess the circumstances about the individuals themselves. Hope-other and dispositional hope are less related probably because hope-other relates more to altruism. The items relate to others and to global circumstances, not to the individuals.

Assessing Burnout Syndrome in Latin-American Priests

Ignacio Pedrosa[1], Helena Lopez Herrera[2], Mª Purificación Vicente Galindo[3], Javier Suárez-Álvarez[1], Mª Purificación Galindo Villardón[3] and Eduardo García-Cueto[1]

[1]University of Oviedo, Oviedo, Spain; [2]UNIBE, Oviedo, Costa Rica; [3]Universidad de Salamanca, Salamanca, Spain

e-mail: npedrosa@cop.es

Burnout syndrome is a common disorder in a wide range of professional groups, being associated with several psychophysiological alterations. Nevertheless, this syndrome has not been analyzed in depth among the clergy. The goal of this presentation is not only to assess the prevalence of burnout syndrome in Latin-American Catholic priests, but also to study the relationship between burnout and both perceived general health and addictive behaviours. In the present study 881 Catholic priests from different Latin-American dioceses were assessed: Mexico, Central America and the Caribbean. The Maslach Burnout Inventory-22, the General Health Questionnaire-28 and the CAGE were applied; we also recorded the priests' rates of cigarette-smoking. The original factor structures of the questionnaires are confirmed in the clerical sample ($\chi^2/df < 3$; GFI>.95; RMSEA <.06), and the reliability is adequate (alpha from .54 to .85). As regards the dimensions of the syndrome as a function of the countries studied, no statistically significant differences were found, except in the exhaustion dimension. Specific cut-off points were set for burnout syndrome

in priests. The authors established the prevalence of the syndrome in this group, with a figure of 25.39%. Furthermore, burnout shows a clear relationship with general health, and may be associated with addiction to substances such as alcohol or tobacco.

Measurement Invariance of Oppositional Defiant Disorder in Spanish Preschoolers: Do ODD Symptoms Mean the Same for Parents and Teachers?

Eva Penelo[1] and Lourdes Ezpeleta[2]

[1]Universitat Autònoma de Barcelona, Bellaterra, Spain; [2]Unitat d'Epidemiologia i de Diagnòstic en Psicopatologia del Desenvolupament, Departament de Psicologia Clínica i de la Salut, Universitat Autònoma de Barcelona, Bellaterra, Spain

e-mail: eva.penelo@uab.cat

Oppositional Defiant Disorder (ODD) is an heterogeneous disorder. Several dimensions of ODD have been identified, but there are not studies about how these dimensions function cross-informant or even cross-sex and whether comparability of ODD means in different groups of responses is guaranteed. To test measurement invariance of ODD symptoms across sex and informant (parents and teachers) in Spanish preschoolers from the general population, according to Burke's model consisting of 3 dimensions: negative affect, oppositional behaviour, and antagonistic behaviour. A community sample of 622 (311 boys and 311 girls) 3 year-old children participated in the study. Their parents and teachers completed an expanded version of the Strength and Difficulties Questionnaire, including the 8 symptoms for ODD referred by DSM-IV-TR. Confirmatory factor analyses for categorical items across sex (multigroup approach) and informant (repeated measures design) were conducted with MPlus7; the multilevel strategy was considered due to cluster sampling for teachers' ratings. Nested models were compared with the scaled chi-square difference (*p* level set at .01). Full measurement and structural invariance across sex within each informant was achieved and latent means did not differ across boys and girls. However, while full metric invariance (equivalence of factor loadings) across informant was obtained, scalar invariance (equivalence of thresholds) was not attained, only 11 of 16 parameters being invariant (68.8%). These differences indicate that equivalence for ratings of parents and teachers is not complete; given the same underlying level of the latent trait, parents tended to rate some symptoms (i.e., "loses temper" and "argues with adults") as more frequent than teachers. ODD appears as a source-specific disorder. The simultaneous use of parents' and teachers' ratings may be considered in the context of a lack of complete measurement invariance, which implies that comparisons of means from both informants are not readily interpretable.

Adapting Spanish Attention Related Driving Error Scale into British English

Elsa Peña Suárez

University of Granada, Granada, Spain

e-mail: elsapsuarez@ugr.es

The Attention Related Driving Errors Scale (ARDES) is a self-reported questionnaire to assess individual differences in the proneness to make attentional errors while driving, which is composed by 19 item-Likerttype. Driver distraction has been widely explored during recent decades, because driver distraction is a contributory factor in road traffic accident. The aim of this study is to develop a British English version from the sourceSpanish ARDES. First, wefollowed a forward-backward-translation design to translate the Spanish ARDES into the targetBritish English. To avoid weaknesses of the translation design, six experts were enrolled in this study divided into two teams. Both teams were composed by a translator, a verifier, which

revised translator's version, and a methodologist, which revised both versions. All were asked to suggest appropriate adaptations of the items, instruction and response scale, according to the British language, culture, traffic regulations and driving habits. The agreement between translator, verifier and methodologist in each team was carried out in a joint meeting. Secondly, psychometric properties of the British English ARDES were examined using responses from British drivers. Psychometrics show adequate values for the British English ARDES measures. Lastly, methodological and practical challenges when adapting into a target context with different culture and driving habits related to the intended construct will be discussed.

Changes in Purpose of and Methods for Setting Cut Scores

Marianne Perie

CETE at the University of Kansas, Lawrence, United States

e-mail: mperie@ku.edu

Over the past 50 years there has been a movement in educational testing in the United States from determining a single cut-off point, typically associated with gaining a certification, to measuring the degree of proficiency along a scale, to now determining readiness for the next step in the educational process. As testing purposes have changed, so have methods for determining whether that performance is "good enough." In the 1960s and 70s, a single cut scores was typically set indicating whether or not a student met a passing mark. Methods included Angoff, Nedelsky, and Ebel. In the 1990s and 2000s, the trend in standards-based testing led to criterion-referenced performance meant to determine how much a student knew. Scales were typically divided into four performance categories: Below Basic, Basic, Proficient, and Advanced with movement across levels valued. The Bookmark method was introduced at this time and became the most popular standard-setting methods for educational tests. Now, with the movement towards measuring "college and career readiness" the field has focused on how best to use assessment results to predict future performance. College entrance exams, the ACT and SAT, have been benchmarked for the first time. State policymakers are asking psychometricians to bring in external validation data to standard setting meetings, and new statistical methods for setting cut scores are emerging. This paper will describe the progression in techniques, provide specific examples from the United States, and focus on the importance of matching the standard setting method to the purpose and use of the assessment.

Testing on Tablets: Creating Economic Disparity?

Marianne Perie[1] and Scott Smith[2]

[1]CETE at the University of Kansas, Lawrence, United States; [2]Kansas Department of Education, Topeka, United States

e-mail: mperie@ku.edu

In a Midwestern state in the USA, over 99% of students take their annual, summative assessment online. In 2013, a pilot test was completed with small population of 4th- and 7th-graders taking the test on an iPad, rather than the traditional laptop or desktop computer. Analysis results showed a slight disadvantage in performance for students taking the test on the iPad. Over the summer, schools purchased new hardware. Some evidence suggests that poorer school districts were more likely to purchase iPads than traditional computers due to cost. In spring of 2014, all students will have the option to take the test on an iPad. This paper will analyze whether there is a systematic difference in performance between the different platforms, the relationship between the platforms and the economic status of the district and school, and

attempt to determine whether this move to iPad could further disadvantage students from poorer economic backgrounds.

Incidences of Items Reverse Scoring on Test Results

Álvaro Postigo, Álvaro Villegas, Javier Suárez-Álvarez and Eduardo García-Cueto
Universidad de Oviedo, Oviedo, Spain
e-mail: alvaro.postigo.gtz@gmail.com

The experts in the development of psychological assessment instruments advise that the positive and negative items be balanced along the scale in order to avoid the acquiescence effect in the variable object of study. This implies that within the same scale, the participant must respond stating their agreement on a set of items and showing his disagreement in the remaining items to get a consistent score with the structure. However, when it comes to correct these scales and obtain a score in the variable, this implies the reversion of those items. This procedure assumes an underlying psychological process, in which if the participant is completely in accordance with a statement, they must totally disagree with the denial of that same item. The aim of this study is to analyze the impact that reversing score has on the outcome of the measurement, taking into account the influence of cognitive and socio-demographics variables. To achieve this objective different socio-demographic, intelligence and self-efficacy scales were applied. The latter scale was presented to all participants three times, containing items written in a positive, negative, and balanced format respectively. The results show that redirection tends to modify the results of measurement. This suffered modification is not independent of the socio-demographic and cognitive aspects of the participants. In conclusion, it seems appropriate to assume that reversing score procedure does not maintain constant measurement levels of the ability of the participant, which presents important practical consequences when carrying out any type of evaluation.

Detecting who is Going to Cause Problems

Achim Preuss, Katharina Lochner and Anja Heins
cut-e Group, Hamburg, Germany
e-mail: achim.preuss@cut-e.com

In employment testing, companies usually screen applicants based on attributes that will allow an employee to be successful (Marcus & Schuler, 2004). However, counterproductive work behaviours such as absence from work, fraud, or dangerous behaviour cause a lot of damage and therefore factors that predict such behaviours should be assessed in employment testing as well (Furnham, Hyde, & Trickey, 2013). A new online instrument predicting counterproductive work behaviours (CWB) was developed, based on the notion that behaviour is best predicted by considering the person and situation (Mischel & Shoda, 1995). Zimbardo (2007) identifies critical aspects of a situation that make counterproductive work behaviours more likely to appear, along with attributes that make a person less susceptible to the cues of the situation. The latter are the facets of the instrument. Sudy 1: Item selection and reliability analysis The data of 331 participants showed that item-total correlations are between .4 and .7 and internal consistencies are around .8. Intercorrelations between the self-ratings on caring, disciplined, and trustworthy and the respective questionnaire facets are between r=.41 ($p < .01$) and .60 ($p<.01$). Study 2: Validity In several single case analyses, questionnaire results were compared to interviews assessing the same facets as the questionnaire. Correlations are between $r = .36$ ($p < .05$) and $r = .67$ ($p < .01$) for the self-ratings by participants and between $r = .37$ ($p < .05$) and $r = .77$ ($p < .01$) for the interviewer ratings. Study 1 demonstrated that the psychometric properties were

satisfactory, Study 2 showed a high congruence between the questionnaire and the single case analyses. However, the instrument needs to be validated with respect to its predictive power of counterproductive behaviour. Studies with addicts and inmates are currently being planned and results will be available early 2014.

Adolescent Pathological Gambling: Using IRT to Construct a Scale Based on the New Gambling Disorder Criteria

Caterina Primi, Francesca Chiesi and Maria Anna Donati

Neurofarba - Section of Psychology University of Florence, Florence, Italy

e-mail: primi@unifi.it

Given the widespread of gambling among adolescents, much attention has been paid to the issue of measurement of youth gambling problems. Nevertheless, there is a debate about the efficiency of the most commonly employed adolescent gambling screens, measures based on the revised diagnostic criteria for Gambling Disorder of the Fifth Edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V, American Psychiatric Association, 2013) are lacking, and there is a general lack of Item Response Theory (IRT) studies on problem gambling severity measures. The aim of this study was to develop a new scale (Gambling Behavior Scale - GBS) to measure pathological gambling behavior. Specifically, we aimed to develop items reflecting the DSM-V diagnostic features and to apply IRT in order to examine the instrument's accuracy along problem gambling severity levels. The validity of the scale was studied with reference to school achievement, quality of life, risk seeking behaviors, and superstitious thinking. The GBS was administered to 249 adolescent past year gamblers (Male = 54%, Mean age=17.55, SD=.50). The unidimensionality of the construct, a fundamental criterion underlying IRT models, was assessed through a Confirmatory Factor Analysis (CFA), IRT was applied to evaluate the functioning of the GBS along the problem gambling severity continuum. Item parameter estimates and the test information function showed that each item and the global scale satisfactorily measured the latent trait. Specifically, items had different levels of severity ranging from medium to high values and the test accurately measured medium-high gambling severity levels. Finally, evidence of the validity of the GBS was provided examining the relationships between problem gambling and the related constructs. These findings provide evidence of the GBS's adequacy for assessment of adolescent problem gambling indicating that it satisfactorily measures different levels of the underlying construct.

Translation and Adaptation of the SOSIE Test for Brazilian Portuguese

Ricardo Primi[1], Paul Mckeown[2], Mireille Simon[3] and Carole Fortier[3]

[1]Universidade São Francisco, Itatiba, Brazil; [2]Pearson Clinical and Talent Assessment, London, United Kingdom; [3]Pearson France - ECPA, Montreuil, France

e-mail: rprimi@mac.com

SOSIE combines the Gordon Personal Profile and Inventory (GGPI), the Survey of Personal Values (SPV) and The Survey of Interpersonal Values (SIV) to access personality traits and values relevant to the recruitment, career management, guidance and training. It uses the forced choice format trying to control the desirability of the responses on subject's choices. This paper shows the translation/adaptation study carried on in Brazil to gather evidence of SOSIE equivalence of the Brazilian version as compared with data from ten countries. Data from 472 persons (managers and college-level professionals across various companies), age M=33.1, SD=9.9, 58.8% women. The adaptation steps were: translation and cultural language adaptation, item analysis, reliability and validity analysis. Positive socially desirable items

endorsements were contrasted against negative ones showing the expected pattern with a correlation of a dummy variable indexing positive item with mean endorsement of .74 (N=124, p<.001). Scales reliability were from .62 to .81 (median of .86). Internal structure validity was assessed by a scale factor analysis resulting in a solution of three behavior styles: leader, organizer and facilitator similar to the configural model found in samples of other countries. The comparison of factor loadings from data of eight countries resulted in high congruency coefficients (median of .83). Convergent validity evidence was assessed through a correlational analysis of SOSIE scales with a big five inventory (BFP). Criterion validity compared persons working at managerial level with a general sample of professionals. All the patterns were coherent with expectations based on international studies of SOSIE. In summary results indicate adaptable equivalence of the Brazilian version of SOSIE.

The Factor Structure of the Spanish Version of the PANAS in Women with Fibromyalgia

Manuel Pulido-Martos[1], Fernando Estévez-López[2], Cristopher J. Armitage[3], Alison Wearden[3], Inmaculada C. Álvarez-Gallardo[2], Víctor Segura-Jiménez[2], Manuel J. Arrayás-Grajera[4], María J. Girela-Rejón[2], Ana Carbonell-Baeza[2], Virginia A. Aparicio[2] and Manuel Delgado-Fernández[2]

[1]University of Jaén, Jaén, Spain; [2]University of Granada, Granada, Spain; [3]University of Manchester, Manchester, United Kingdom; [4]University of Huelva, Huelva, Spain

e-mail: mpulido@ujaen.es

Fibromyalgia is a clinical syndrome characterized by the presence of widespread pain, hyperalgesia, and allodynia. People with fibromyalgia report lower levels of positive affect and higher levels of negative affect than healthy peers. However, the Positive and Negative Affect Schedule (PANAS) that is frequently used to assess affect is sensitive to sampling and its factorial structure is controversial. The internal structure of a measurement instrument has an impact on the validity of the score interpretations. To assess the factor structure of the Spanish version of the PANAS in women with fibromyalgia. The design of the study was cross-sectiona[l]. Using confirmatory factor analysis (CFA), data from a sample of 442 adult women with fibromyalgia was analyzed. Initial analyses tested the fit of the present dataset against the original two uncorrelated factor structure (positive affect, negative affect) but this model was rejected. The best fit was obtained with a two correlated factor structure where the error terms of some items were allowed to correlate. It suggested that interpretation of the PANAS scores from people with fibromyalgia population need to be based on the two correlated factor structure emerged from our data. Further research among other samples of people with chronic pain diseases is required to further understand the structure of affect.

Adaptation and Validation of Connor-David Resilience Scale (CD-RISC) in the Spanish Population

Manuel Pulido-Martos, Mariola Fernández and Esther Lopez-Zafra

University of Jaén, Jaén, Spain

e-mail: mpulido@ujaen.es

Resilience research has become a very important field of study. Resilience involves coping ability, adaptation to the environment and trauma recovery, which improves quality of life. The number of studies in recent years has increased considerably due to the development of instruments that measure and evaluate this construct. The most commonly used and internationally recognised is the Connor-Davidson Resilience Scale (CD-RISC; Connor and Davidson, 2003). Bobes et al. (2008) translated this scale into Spanish but no study has addressed the validity and psychometric properties of this scale in the Spanish population. Our main objective is to adapt

this scale and to analyse the reliability and validity of the Spanish version of the CD-RISC. First, we followed the ITC guidelines to make the traduction and adaptation of the CD-RISC into Spanish by the back-translation steps. Second, a heterogenic and representative of the population sample of 972 individuals (503 women and 470 men) with a mean age of 31.87 (SD = 10.45, range = 18-56) completed a questionnaire that comprised the translated and adapted CD-RISC along with two other questionnaires to measure alexitimia and cope strategies. We tested for models for the questionnaire structure. Model A: five-factor model as the original structure. Model B: four factors yielded in other countries studies. Model C: three factors based on our exploratory analysis and Model D, which included two factors. Our results showed that Model D obtained better parameters. Therefore, a 21-item version that includes the basic elements of resilience and has adequate reliability is proposed. We also analysed the relationship of the proposed scale to other criterion variables. This Spanish version of the CD-RISC was found to be a good tool for analysing resilience in Spanish samples and for use in cross-cultural comparisons.

Validating a Test of English (SET): A Multicompetence, Structural Equation Modeling Approach

Kioumars Razavipour

Shahid Chamran University, Ahvaz, Iran

e-mail: razavipur57@gmail.com

The Specialized English Test (SET) is an English test used for screening candidates who apply to English language programs at Iranian state universities. A range of invalidity evidence has been collected blaming the test for the construct-irrelevant variance it introduces to test scores. Previous validation studies have almost all approached the test from the vantage point of conventional communicative competence theories (see Bachman & palmer, 2010). We speculated that a different underlying theory, the multicompetence theory, might be more plausible model to examine the validity of a test that comprises items from three different languages, namely English, Persian, and Arabic. To this aim, we collected the test scores from two hundred test takers, selected through a stratified sampling approach, and analyzed them using structural equation modeling (SEM). Results revealed a satisfactory model fit between the data and the multicompetence theory we proposed. In addition to providing validity evidence for the test, the findings clearly indicate that in test validation the underlying theory which informs the process is of primary importance, attesting to Messick's assertion that theoretical argument coupled with empirical evidence should be used in investigating the validity of the inferences that are to made of test scores. The implications for language assessment, particularly in multicultural and multilingual settings are further discussed.

Translating and Adapting Psychological Tests for French-Canadians: A review of methods Used Since 2000

François René De Cotret and Francoeur Aline

Université Laval, Quebec, Canada

e-mail: francois.rene-de-cotret.1@ulaval.ca

With two official languages—English and French—and a province—Quebec—where 85% of the population speaks French, Canada is a country in which translating psychological tests is a necessary practice. Yet, no review of the literature has been undertaken to establish how this practice is pursued, and if guidelines such as those presented by the ITC back in 1999, and widely

disseminated since then, are applied. Our study aims at finding out what range of methods have been used to translate/adapt tests for the French-Canadian population since the presentation of the ITC's Guidelines, and if some methods and principles are more popular than others. We have identified 50 research articles published between 2000 and 2013 that describe the validation process when it comes to psychological tests for the French-Canadian population. The information provided by the authors about the translation process per se has also been collected for further analysis. So far, an analysis of 20 of the 50 articles shows that the translation methods used vary greatly, that back translation is still among the most popular routes translation takes, and that only a minority of researchers seems to include qualified translators or linguists in the process. Furthermore, the ITC's Guidelines do not seem to have had much impact on the way translation/adaptation has been done in the Canadian context. When fully completed by the end of May 2014, our study, we feel, will surely contribute to the general understanding of the challenging situation the translation and adaptation of psychological tests represent especially today.

Using the Dispositional Flow Scale – 2 to Identify the Key Components of Flow

Scott Ross[1] and Heidi Keiser[2]

[1]DePauw University, Greencastle, United States; [2]University of Minnesota - Twin Cities, Minneapolis, United States

e-mail: srross@depauw.edu

Csikszentmihalyi (1975) first coined the term "flow" to describe 'peak experiences' in a wide array of activities. Phenomenological studies have highlighted key components of flow, which include distorted passage of time, loss of self-consciousness, intense concentration, clear goals and feedback, a balance of challenge and skill, merging of actions and awareness, a sense of control, and enjoyment in flow states. Collectively, these components have good consensus among researchers in describing the flow experience. However, an issue of debate among flow researchers is which components of flow are most important in characterizing and possibly generating the flow state. In 293 young adults, we administered the Dispositional Flow Scale – 2 (DFS-2; Jackson & Eklund, 2002) targeted to general life. The DFS-2 is one of the few psychometrically promising measures of flow, with demonstrated factorial validity (see Eklund & Jackson, 2002), and criterion-related validity (see Ross & Keiser, 2014; Johnson, Keiser, Skarin, & Ross, in press). Thirty-six items comprise nine subscales. Each subscale is composed of 4 items each on a 4-point Likert scale. Using principal components analysis, we identified one item (either the second- or first-highest loading) from each DFS-2 subscale for representation of that flow component in a 9-item composite flow index (alpha = .75). Only the item representing time distortion had a corrected item-total correlation below .30. To determine the components of flow that most contribute to the prediction of our composite flow index, we used multiple regression in which each subscale was recomputed by removing the item used in deriving our composite flow index. Merging of Action and Awareness, Concentration, and Autotelic Experience (enjoyment) were found to be the best predictors of global flow. When backward entry was used, both Time Transformation and Challenge-Skill Balance were removed. Implications of these findings for flow research will be discussed.

Application of Addenbrooke's Cognitive Examination – Revised for Differential Diagnostics of Dementia and Depression

Augustinas Rotomskis[1], Albinas Bagdonas[1], Arunas Germanavicius[1], Neringa Grigutyte[1] and Ramune Margeviciute[2]

[1]Vilnius University, Vilnius, Lithuania; [2]The University of Edinburg, Edinburg, United Kingdom

e-mail: rotomskis.augustinas@gmail.com

One of the usual problems psychologists face in clinical practice is differential diagnostics of dementia and depression. Up to 32 % of old people with depression are misdiagnosed as possibly having dementia (Crigger and Forbes, 1997; Marin et al., 2002). It has been reported that the Addenbrooke Cognitive Examination – Revised (ACE-R) could discriminate the cognitive decline due to depression from that due to dementia (Dudas et al., 2005), although this is not uniform in all cultural backgrounds (Stokholm et al., 2009). The aim of the study was to investigate, whether the use of ACE-R differentiates depression from dementia in a Lithuanian-speaking population. All participants were older than 50 years. A total sample of 450 individuals were enrolled: 150 with severe dementia, 150 with early stage dementia, and 150 healthy controls without any emotional or cognitive disturbances. It was found that participants with severe depression performed worse than participants without emotional or cognitive disturbances, but better than participants with dementia. Participants with severe depression were differentiated as having lower scores in memory and verbal fluency domains. The ACE-R is a suitable tool to use in clinical practice for differentiation between dementia and depression. Mild impairment in the total ACE-R scores, along with a low score on the memory tasks and verbal fluency, are strongly indicative of an affective, as opposed to organic, pathology. A total score below the cutoff <74 is strongly predictive of an underlying dementia.

Exploring the Accuracy and Reliability of Angoff Cut Score Judgments

Paul Sackett[1] and John Weiner[2]

[1]University of Minnesota-Twin Cities, Minneapolis, United States; [2]PSI, Burbank, United States

e-mail: psackett@umn.edu

Despite the variety of approaches that have been proposed for setting cut scores, the "modified Angoff" approach (Angoff, 1971) appears to be the most widely used. The current work examines Angoff judgments in the context of licensure and certification programs. The data presented come from a testing program for a national Real Estate Sales licensure examination. Judgments were made for 106 items by 19 subject matter experts (SMEs), and these SME judgments were compared with subsequent item performance (e.g., percent correct) with a median N of about 50,000 examinees per item. The first question addressed was the degree of correspondence between SME judgments and item p-values. In essence this sheds light on whether SMEs can rank the difficulty of items accurately. Across the 106 items, the median correlation between individual initial SME ratings and p-values was .31, with a range of .21-.48. Thus, while some raters are modestly more effective than others in differentiating easy from hard items, none prove highly effective at doing so. The mean ratings across the 19 raters showed modest correspondence with p-values (r = .48). Shifting to the second round of ratings, where p-values have been shared with raters, the median r between ratings and p-values increased from .31 to .71. The mean across raters correlated .83 with p-values. Thus, not surprisingly, judgments about the performance of minimally qualified candidates are heavily influenced by information about item p-values. A second research question involved the reliability of SME judgments. These results show that a) Round 2 ratings are more reliable than initial ratings, b) Round 2 reliabilities greater than .80 can be achieved with 5 raters, and c) Round 2 reliabilities greater

than .90 can be achieved with 10 raters. This contributes to our understanding of the number of SMEs needed for the Angoff process.

MATRICES, a New Test for Assessing General Cognitive Ability: Development and Results

Pablo Santamaría[1], Fernando Sánchez-Sánchez[1] and Francisco J. Abad[2]

[1]TEA Ediciones, Madrid, Spain; [2]Universidad Autónoma de Madrid, Madrid, Spain

e-mail: pablo.santamaria@teaediciones.com

Matrices items are a popular measure of higher order general cognitive ability (g), being the Raven Progressive Matrices the most widely used and recognized test in the area. However, some major drawbacks may curtail its use nowadays (v.g., widespread items with high exposure rates on the Internet, old norms in certain countries, administration time...). MATRICES test was developed in order to provide a more efficient measure of general cognitive ability (g) using Matrix item. MATRICES test was developed to provide a measure with (1) large, representative and updated standardization samples; (2) different paper and pencil forms designed to different range of ages in order to improve accuracy of the measure with shorter administration time (3) more attractive item design and (4) an optional and alternative CAT administration. A sample of 12.280 individuals from 5 to 74 years old was assessed. The sample selection method took into account the data provided by the Spanish Census and other official information regarding the composition of the Spanish population in variables such as gender, geographical region or educational level in order to guarantee the representativeness of the sample. Multiple norms were developed for different age ranges and forms by using the continuous norming procedure. Internal consistency and test-retest were obtained as indicators of the reliability of MATRICES scores. IRT and confirmatory factor analysis was carried out to study the structure underlying the instrument. Correlations with other measures of intelligence were also explored. Finally, a variety of samples of children and adults with intellectual disability were assessed. The satisfactory results show that MATRICES is a reliable and valid measure that allows assessing general cognitive ability throughout lifespan.

Cognitive Abilities and Gender. A Measurement Invariance Study

Eneko Sarasua Garcia[1], Josu Mujika[1], Paula Elosua[1] and Leandro S. Almeida[2]

[1]Euskal Herriko Unibertsitatea EHU/UPV, Donostia, Spain; [2]University of Minho, Braga, Portugal

e-mail: esarasua001@ikasle.ehu.es

Gender differences in cognitive abilities are a controversial and complex topic. It has been addressed from different areas of psychology, neuropsychology, development psychology, anthropology or ethology to mention some examples. In this controversial context the aim of this work was to offer a psychometric approach to the study of gender differences. The approach is based on the concept of measurement invariance. In order to do that, a general intelligence test battery called Arrazoiketa-Proba Andana (APA; Battery of Reasoning Tests) was applied to a large sample of students (N = 3326) and measurement invariance across gender was assessed. The battery evaluates cognitive skills and general intelligence (g) of students aged between 9 and 18 years. According to the age APA is available in three forms: APA-1 (9-12 years old) consists in abstract, verbal, numerical and practical reasoning tests; APA2 (12-15) and APA3 (15-18) and are composed by abstract, verbal, spatial, numerical and mechanical reasoning tests. A progressive assessment of factorial invariance that involved analyses over configural, metric and strong equivalence was conducted on each one of the forms of APA. Partial invariance was concluded. For verbal and numerical reasoning tests intercepts were higher for females in

APA-1 and APA-2. Differences in the intercept parameter were found in the mechanical reasoning test in APA-3.

A Mixed Approach to Data Cleaning of a Large-Scale High-Stakes Assessment Test

Michele Settanni[1], Renato Miceli[2] and Davide Marengo[1]
[1]University of Torino, Torino, Italy; [2]University of Aosta Valley, Aosta, Italy
e-mail: michele.settanni@unito.it

Aberrant response patterns resulting from the involvement in cheating behavior is particularly problematic for the assessment of test validity and reliability. Besides cheating, other factors can have a significant impact on the test variability, posing a threat to the test evaluation (e.g. test anxiety, low motivation, teacher disaffection with the test). These factors can result in aberrant response behavior (ARB) which are not easily identifiable. Several approaches to detect ARB have been proposed, some of them based on the examination of contextual information other than the test performance. Concerning methods which do not rely on contextual information, three different approaches can be identified: 1. Comparison of individual response patterns; 2. Fit of response patterns to a measurement model; 3. Model-free statistics. None of these methods have reached a consensus among scholars. Aim of this paper is to show results of the application of a composite model to identify ARB, based on multiple indicators, considering the influence of both individual and aggregated class-level indicators. Data provided by Italian National Evaluation Institute (INVALSI) comprised the student population which undertook the INVALSI 2012 standardized math test (N= 519003). Dataset was inspected to identify problematic data according to a two-step procedure. The procedure is based on both a model-free approach and on response data fit to a measurement (Rasch) model. The applied procedure permitted to identify a relevant percentage of "suspicious" data (9.4%). Their removal from the working data set did not significantly affect the item parameter estimation. We presented a procedure appropriate for cleaning data from individual aberrant response patterns in situations in which no contextual variables are disposable. Strengths of this procedure are twofold: 1. Retaining only individuals with non-ARB, aggregated level (classrooms, schools, regions) statistics are more reliable and 2. Obtaining trustworthy item parameter estimates.

The Predictive Power of Customer-Focused Sales Competencies

Levent Sevinç and Yasin Rofcanin
Assessment Systems Turkey, Istanbul, Turkey
e-mail: levents@assessment.com.tr

Recent research has shown that competent employees are critical resources for companies to grow competitive business settings (Redmond, 2013). Given that most of the businesses have become service-oriented especially in the last two decades (Batt, 2002; Ford & Bowen, 2008), development of certain competencies has become one of the strategic priorities for employers. Building on the recent calls for research, we carried out this study with the purpose of demonstrating the critical role of customer-focused competencies in predicting performance outcomes. We conducted a field study (N = 103) in one of the leading banks in Istanbul, Turkey. Participants were selected from departments where sales function was central. We utilized five customer focused sales competencies (i.e., persuasion, relationship ability, effective communication, success orientation and customer orientation). Findings showed that relationship ability (R^2=.15; $p<.05$) and persuasion (R^2=.13; $p<.05$) had the highest degree of explanatory power of performance (R^2=.07, $p<.001$). To see if there was statistically significant difference among competency items, we constrained all the paths to be equal over performance

(Byrne, 2001; Kline, 1998). Model fit decreased which is an indication of unique explanatory power of each of the competency dimensions (? df = 1; ? □²= 3.85, p < .05). Drawing from the findings of this study, we hope to contribute to competency literature in two ways. Our first contribution relates to the measurement of a specific sales-related measure, which has been overlooked in the recent studies. With the unprecedented growth of service industries, strategic human resources practices should focus on the development of customer-oriented sales competencies. Our second contributes lies in associating our specific measure to objective performance, which is critical for employers. We, therefore, hope to bring a new perspective on sales competencies within an emerging economy.

Translation and Adaptation of the Middle School Mathematics and the Institutional Setting of Teaching (MIST) Teacher Survey into Turkish Language and Culture

Sevim Sevgi[1], Giray Berberoğlu[1], Paul Cobb[2] and Thomas M. Smith[2]

[1]Middle East Technical University, Ankara, Turkey; [2]Vanderbilt University, Nashville /TN, United States

e-mail: sevimsevgi@gmail.com

The purpose of this initial study is to translate and adapt the Middle School Mathematics and the Institutional Setting of Teaching (MIST) Teacher Survey for Turkish teachers. The MIST Teacher Survey was originally designed by Cobb and Smith (2008) to assess teachers' perceptions of the support structures for developing ambitious instructional practices in mathematics.The purpose of adapting the scale into the Turkish language culture and education system is two-fold. First, the study will investigate how the scale functions similarly across the different cultural contexts and languages. Second, Turkish and American teachers will be compared in the dimensions of self efficacy, school climate, teacher teacher trust; and teacher deprivatization of practice. The scale was translated from English into Turkish language by three independent translators. The researchers evaluated the translation fidelity of the three versions, and prepared a single version of the Turkish instrument by selecting the items which best reflected the meaning of the original statement. After revisions, for further validation of the translated version of the scale, the instrument was administered in a group of six middle school Turkish teachers as a cognitive interview. As a result of the interview, a further revision was carried out in order to avoid possible sources of ambiguity in the item content. The final version of the scale was administered in a group of 134 Turkish teachers. The American sample consists of 259 teachers. Across these groups the factor structure equalities will be carried out for the selected sub-dimensions of MIST. In the final step, the means of the two samples will be compared across the sub-dimensions by multivariate analysis of variance design. The differences and similarities of the teachers in two different languages, cultures and education system will be evaluated and discussed in the paper.

Notes on Measuring Cognitive Complexity of Shapes Based on Information Theory

Shen-Guan Shih

Department of Architecture, National Taiwan University of Science and Technology, Taipei, Taiwan

e-mail: soreros@gmail.com

Mental rotation, recognition, puzzle, tiling, dissection of shapes are various kinds of geometric problems that are often used in testing the ability of spatial cognition. One of the most fundamental problems of such testing is regarding the complexity of shapes from the perspective of human cognition. Complexity can be measured by the quantity of information that is required for the resolution of problems. This paper discusses issues regarding the measurement of shape complexity based on the theory of information, within which information is a measurable quantity of messages. Spatial cognition can be classified as a kind of communication systems, of

which shapes in space are information sources that continuously send message to human mind through signals that are encoded as sights and touches by the sensory system. The recognition of shapes relies on the ability of decoding such visual and tangible signal into messages that can be processed and reasoned by the cognitive system. In Shannon's mathematical model (Shannon 1949), information, as a synonym of entropy, is defined as the measurement of uncertainty that is disclosed by the received message. The hypothesis of this research is that as the observer's eyes take sights on parts of interest within a shape, low level geometric features are recognized and processed to derive the higher structure of the shape, in a way similar to letters and words that are read to uncover the syntactical structure and eventually the semantics of sentences and paragraphs. Based on the hypothesis, Shannon's method for measuring information sources that generate textual messages is adapted to measure the cognitive complexity of shapes. It is expected that the finding of this research would facilitate ability testing of spatial cognition. Methods for the formulation of complexity measurement on geometric shapes are investigated and discussed.

Validating the Inferences Made from the 2012 Mathematics PISA

Pooja Shivraj

Southern Methodist University, Dallas, United States

e-mail: pshivraj@smu.edu

The Programme of International Student Assessment (PISA) has been administered to 15-year olds every three years since 2000. Since then, the US has performed below the OECD mathematics average, and way below countries such as Finland, China, Japan, Korea, and Singapore. The objective of this paper is to study whether fairness could be an explanatory factor in the poor rankings of the US on the PISA. It is investigated using two separate definitions of fairness, with supporting pieces of evidence for each hypothesis: (1) The mathematics assessment framework developed for the 2012 PISA was created using the content standards of eleven countries, none of which was the US. Lack of US' opportunity to learn (OTL) the content assessed is investigated as the first aspect of fairness; and (2) PISA mathematics scores are more correlated with PISA reading scores than content knowledge, questioning whether the construct and items are biased. Bias on the items is investigated as the other aspect of fairness. Kane's validity framework is used to design the study. OTL and lack of bias are investigated as assumptions underlying the inferences made from the PISA scores, and an alignment study and a differential item functioning analysis are the pieces of validity evidence used to support these assumptions. The alignment study aligns the PISA framework and items to the enacted curriculum from middle school to tenth grade to determine if US students have the OTL the PISA content. A logistic regression is conducted to determine if items function differently for US students versus students in high-performing countries whose (1) curricula developed the PISA framework, and (2) primary language is English. A qualitative review of the items is done to determine source of DIF. This dissertation data is being analyzed.

Development and Validation of a New Test to Measure Emotional Intelligence in the Workplace

Katja Schlegel, Marcello Mortillaro and Irene Rotondi
Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland
e-mail: katja.schlegel@unige.ch

Emotional intelligence (EI) has been one of the most widely debated psychological constructs in the past decade. Specifically, the debate centers around how EI should be defined and how it should be measured. The „trait EI" approach defines EI as a variety of non-cognitive traits and dispositions that are linked to success in coping with environmental demands and that can be measured with self-report questionnaires. In contrast, the „ability EI" approach considers EI as a cognitive ability that can be measured with performance-based tests. Given that trait EI measures have been shown to overlap highly with existing measures of personality, ability EI is to date considered the more appropriate conceptualization to be labeled EI. However, the one widely used test developed within this framework, the Mayer-Salovey-Caruso EI Test (MSCEIT), has been criticized for its psychometric quality and for its scoring based on group consensus. Here, we introduce a new performance-based EI test targeted to employees that includes subtests for emotion recognition, emotion understanding, emotion management, and emotion regulation. The emotion recognition subtest consists of short video clips in which actors express 14 different emotions. For the other three subtests, items were developed based on the Situational Judgment Test approach. Real workplace scenarios were collected through interviews with employees and managers and the response options (including the correct answer) were derived from various emotion theories instead of group consensus. The psychometric properties of this new test were examined in a study with 155 students. Results showed that the internal consistency and the difficulty level of each subtest were acceptable. In addition, construct validity was established through meaningful correlations with other standardized tests and questionnaires for each subcomponent of the test. Given these promising results, future studies should investigate the predictive validity of the test with respect to workplace performance.

Validating Psychological Tests: Trends and International Differences. A Bibliometric Study

Jennifer Schroth, Saskia Naescher, Günter Krampen and Gabriel Schui
ZPID, Trier, Germany
e-mail: schroth@zpid.de

Professional psychological testing and assessment is only possible when existing scientific standards and guidelines on test construction and test use are considered. Well-established examples are the Standards for Educational and Psychological Testing (AERA, APA, & CME, 1999), the International Guidelines for Test Use (ITC, 2001), or the German DIN norm 33430 (Kersting, 2008). In addition to several important aspects, the continuous examination of test validity represents one of the essential issues in all these standards. For the purpose of investigating international differences in the amount and development of validity studies, literature databases of several countries were bibliometrically examined. Present analyses were based on entries in PsycINFO (US), PSYNDEX (Germany), PASCAL (France), ISOC-Psicologia (Spain), MEDLINE US), and ERIC (US). With the exception of PsycINFO, all databases were accessed through PubPsych, an open access retrieval system for psychological resources maintained by the Leibniz Institute of Psychology information (ZPID). In each database, specific retrieval criteria were used to identify journal articles with a main focus on test validation and a publication date between 1983 and 2013. The results show increasing quantities of validity studies in almost all databases. A more detailed look at the data reveals interesting distinctions in

respect to their trends, depending on the thematic emphases of the particular databases. Time series modeling predicts a continued development in this direction. The focus of current research appears to be less on the construction of new instruments but rather on the further validation of existing ones, exemplified by the well-documented decrease of new psychological tests published in the German-speaking countries.

Examining Cultural Validity: Self-Esteem Across Social Contexts

Christina Simmons[1] and Pedro Portes[2]

[1]University of Georgia, Athens, United States; [2]Center for Latino Achievement and Success in Education (CLASE); University of Georgia, Athens, United States

e-mail: simmons@uga.edu

Establishing cultural validity for psychological measures is well-recognized in cross-cultural research; however, often instruments are translated and drawn without considering cultural differences. The Rosenberg Self-Esteem scale (RSE) is an example of a measure applied across cultural groups in 53 nations. Different ethnic groups show differences in self-esteem; however, generalizations appear flawed as the construct itself may differ by each culture's valued traits. This study explores cultural validity of one psychological measure to address broader validity issues in the field. We examine consistency of RSE factor structure(s) based on 'ideal self' of Americans and Japanese. If ideal self for Japanese is not represented, the scale would lack validity, justifying further exploration. Participants include Caucasian Americans (n=234; M=20.27yrs) and Japanese (n=311; M=19.62yrs). Multivariate analyses, including exploratory and confirmatory factor analyses, compare Americans and Japanese on the RSE, Social Desirability scale(SD) and Social Comparison scale(SC). Correlations are calculated between measures. Significant correlations were found between the SD and SC for Americans and Japanese (r=.50and r=.45, respectively) and between the RSE and the SD, albeit modest (r =37and r =35, respectively). Ethnicity was significantly related to RSE items, Pillai's Trace=.836,F(5,526)=537.54. The strength of association is partial eta^2=.836. Americans scored higher than Japanese on all scales. The RSE as a unidimensional construct was justified for Americans; all items had positive factor loadings over .60. However, Japanese self-esteem was characterized by low factor loadings and items with negative values. Thus, the RSE's semantic structure differs for Americans and Japanese. Our findings suggest that the RSE and SC lack cultural validity in spite of marginal factorial structure fit. Measuring self-esteem and other constructs with measures based on Western individualistic values may lead to misleading conclusions. We propose a sequence of steps to establish cultural validity using these and other cross-cultural data.

The Utility Gain of Leaving Professional Judgement Out of Prediction: Clinical Versus Mechanical Interpretation of GMA and Personality

Sofia Sjöberg

Pearson, Stockholm university, Stockholm, Sweden

e-mail: sofia.sjoberg@psychology.su.se

The purpose of this study was to analyze and illustrate the margin utility of using clinical versus mechanical data combination for personnel selection purposes using measures of personality and general mental ability as predictors of job performance. By utilizing meta-analytic estimates for personality and general mental ability to predict job performance, and for clinical versus mechanical data combination predicting work criteria, utility analysis was applied to estimate the margin utility between data combination approach for different selection scenarios. The findings

indicate that in a selection context, the difference in financial outcome is likely to be extensive between the two data combination methods. The gain in utility of combining data mechanically corresponds to an amount likely to represent the difference between failure and success for many organizations. This comparison provide professionals with the opportunity to gain insight into the difference in financial outcome of applying data combination method and by that increase the likelihood of acceptance and use of the mechanical approach. It also provides the reader with an example of how to utilize estimates provided by research, how to apply them for data combination purposes, and how to estimate the margin utility in their own selection practice.

Pilot Selection in the Swedish Air Force: Preliminary results from an Incremental Validity Study of Cognitive Ability and Interviews

Anders Sjöberg[1] and Wolgers Gerhard[2]

[1]Stocholm university, Stockholm, Sweden; [2]The Swedish Armed Forces, Stockholm, Sweden

e-mail: anders.sjoberg@psychology.su.se

Historically, selection of military pilots has been an area of great research effort due to the enormous costs of candidates failing in pilot training (Hunter & Burke, 2002). Research supports the use of different types of cognitive tests to predict the future performance of pilots (Martinussen & Torjussen, 1998). In practice however, cognitive test are rarely used in isolation. Despite scarce research establishing incremental validity of using interview data in addition to cognitive tests in pilot selection, interviews are commonly conducted before the selection decision is made. The overall purpose of this study is to evaluate the incremental validity of a quasi-structured interview, currently used for pilot selection to the Swedish Air Force, over and above a cognitive test battery. This was accomplished by estimating the relationship between the interview score and the total score of a cognitive battery by using data from the Swedish Pilot Selection Database (SPSD, N=425). In addition, a small-scale meta-analysis (k=2; N=504) of primary validation studies was conducted to estimate the relationship between the predictor scores (i.e., cognitive ability and interview) against pilot performance collected during the basic training period. Finally, a stepwise regression analyses were conducted to answer the question about incremental validity. Results show that the interview score add incremental validity over the cognitive ability score, and cognitive ability score add incremental validity over the interview score. The use of both cognitive tests and interviews in future pilot selection are discussed from a utility perspective.

Analysis of Covariates Using TIMSS Data Based on Multiple-Groups Higher-Order Reparameterized DINA Model

Yoon Soo Park[1] and Young-Sun Lee[2]

[1]University of Illinois at Chicago, Chicago, IL, United States; [2]Teachers College, Columbia University, New York, NY, United States

e-mail: yspark2@uic.edu

A cognitive diagnostic model for multiple groups that allows the specification of covariates affecting attribute classification using the higher-order reparameterized deterministic input noisy "and" gate (HO-RDINA) model is proposed. Here, the probability of solving an item correctly, given a binary latent variable that indicates whether an examinee has mastered the complete set of attributes required for answering an item, is expressed in terms of false alarm and detection parameters; attributes, in turn, become indicators of a higher-order latent trait. In the Multiple-Groups RDINA (MG-RDINA) model, the distribution of attributes may differ between

groups, and a covariate can be specified to affect attribute classification. This study examines the utility of the covariate multiple-groups HO-RDINA (covariate MGHO-RDINA) model using the Trends in International Mathematics and Science Study (TIMSS) data. Data from TIMSS 2007 4th grade mathematics were used, measuring 5 attributes: (1) whole numbers, (2) fractions/decimals and number patterns/relationships, (3) lines/angles and shapes, (4) location/movement, and (5) data display. Six participating countries were used as groups in the covariate MGHO-RDINA model: Germany (Rank #14, n=362), New Zealand (Rank #29, n=345), Qatar (Rank #42, n=507), Russia (Rank #8, n=326), Singapore (Rank #2, n=356), and USA (Rank #13, n=565). Data were fit using LatentGOLD. Gender and science ability were specified as covariates. The covariate MGHO-RDINA model had the best model fit based on BIC indices (RDINA=68,724; covariate RDINA=67,584; MG-RDINA=66,056; covariate MG-RDINA =64,943; MGHO-RDINA=65,667; covariate MGHO-RDINA=64,941). Results of the covariate parameters showed differences in effects by country, particularly the effect of science ability on attribute mastery, indicating utility of the model. The framework proposed for examining the effect of covariates on attribute-level parameters has various applications in large-scale assessments that have multiple groups. The proposed covariate MGHO-RDINA model can provide fine-grained information on attribute classification and their relationship with covariates.

A New Scale for Assessing Optimism in Youth

Javier Suárez-Álvarez, Ignacio Pedrosa, José Muñiz and Eduardo García-Cueto

Universidad de Oviedo, Oviedo, Spain

e-mail: suarezajavier@uniovi.es

The positive psychology approach has reached a relevant position in modern psychology. From this approach, the optimism has been shown as a key variable in the prediction of personal well-being on several populations and professional contexts. However, not many measurement instruments were developed to assess optimism, and most of them show important psychometric limitations. The main purpose of this research was the development of a new scale for assessing optimism in adolescent population, as well as the study of validity evidence of the test developed in relation to the Big Five and Emotional Intelligence. The sample was formed by 2,693 Spanish adolescents (M=16.52, SD=1.38), from whom the 51,10% were males. In addition to the optimism scale developed, the Overall Personality Assessment Scale (OPERAS) and the Trait Meta-Mood Scale-24 was applied. The optimism scale showed a high reliability (α=.84) and an accurate measurement for a large range of ability (□entre -3 y +1). The factorial structure was essentially one-dimensional. Finally, a high correlation between optimism scores and emotional stability (rxy=.62), as well as between optimism and emotional repair (rxy=.62) were found. The new scale developed for assessing optimism in adolescent population shows high reliability, adequate factorial validity and high convergent validity with stability and emotional repair.

Development the Wechsler Adult Intelligence Scale – IV (WAIS-IV) for the Indonesian Population: A Preliminary Study

Christiany Suwartono[1], Magdalena Halim[1], Lidia Hidajat[1], Marc Hendriks[2] and Roy Kessels[2]

[1]Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia; [2]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Netherlands

e-mail: csuwartono@gmail.com

From lots of tools, the most used and applied is the psychological measurement of cognitive functioning, especially intelligence. Unfortunately, the most frequent used for intelligence

measurement in Indonesia is the Wechsler-Bellevue Adult Intelligence Scale (WB) which was developed in 1939 and thus very out of date. Internationally, WB has been extended and modified several times before becoming The Wechsler Adult Intelligence Test – Fourth Edition (WAIS-IV). Through the years, several Wechsler's measurement tools, such as WB and WAIS are used, in clinical, educational or industrial settings in Indonesia. However, all these instruments are only translated decades ago and not standardized. The primary aim of this study was to construct a translated Wechsler Adult Intelligence Test-IV in Indonesian. Here we describe the first phase in the adaptation of the WAIS-IV in the Indonesian language; including translation, item analysis, and reliability of several subtests. Using convenience sampling, we had 176 participants who have various age, educational background, and ethnicities. Half of the participants (50.5%) were men and 49.5% were women. Age, ranged from 16.2 – 86.1 years old (M = 39.18, SD = 20.39). We conclude that the item sequence of the US version WAIS-IV cannot be applied in Indonesia due to its fluctuation of index difficulties. As a result we rearranged the items sequence in 15 subtests. The subtest reliability coefficients range from acceptable (BD, SI, and CO) to excellent (LN and FW). For the subtests from Verbal Comprehension index (SI, VC, and CO), the interrater agreement was high. So, adaptation of the WAIS-IV is psychometrically promising. Then, next step is to explore the factorial structure of the Indonesian WAIS–IV subtests, to test whether it measure the same of structure of Cattell-Horn-Carroll (CHC) theory.

Exploratory Factor Analysis of the Indonesian Wechsler Adult Intelligence Scale – 4th Edition (WAIS-IV)

Christiany Suwartono[1], Marc Hendriks[2], Weny Sembiring[1], Magdalena Halim[1] and Roy Kessels[2]

[1]Universitas Katolik Indonesia Atma Jaya, Jakarta Selatan, Indonesia; [2]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Netherlands

e-mail: csuwartono@gmail.com

The development of Indonesian version of WAIS IV has just begun. The primary result from our first pilot study (Suwartono, et al) seemed to be psychometrically promising. Therefore in this study we examined the factorial structure of Indonesian translated version of WAIS-IV in Indonesian sample. We had 444 participants, whose data can be analyzed. Most of the participants were female (59%) with age varied between 16 - 69 years old (M = 33, SD = 13.57). They belong to various educational backgrounds, from junior high school until doctoral degree. We used exploratory factor analysis with principal components extraction Varimax Rotation and eigenvalues more than one. The results we got two components extracted. The two components named Perceptual Performance and Verbal Comprehension. Future research can confirm this result with confirmatory factor analysis.

Assess Dimensionality in Order to Optimize Design and Scores

Eileen Talento-Miller, Kyung (Chris) Han, Fanmin Guo and Lawrence Rudner

Graduate Management Admission Council, Reston, United States

e-mail: etalento-miller@gmac.com

To meet the challenge of providing evidence that a measure is valid, psychometricians must document that the scores reflect variations in an unobservable attribute such as reasoning or employability. One of the most important forms of such construct evidence is the agreement between the underlying statistical structure of the test and the intended psychological or cognitive structure. The objective of this session is for attendees to better understand techniques and options for assessing dimensionality. Such information is crucial for optimizing test design

and developing section and total scores. Emphasis of the session is on concepts and generalizability to other examination programs. The presentation will illustrate practical applications of statistical approaches using multidimensional data from a new admissions examination. The example will also address the very real challenges of conducting dimensionality analyses using data from a computer adaptive administration, including dealing with data that cannot be considered missing at random. The first approach demonstrated will be factor analysis, which is the traditional approach to evaluating dimensionality. The philosophies behind the use of both confirmatory and exploratory factor analysis and the options for their implementation will be presented along with the results with this data. Structural equation modeling (SEM) will be addressed next, as it has been a popular choice in recent years when a test consists of multiple, complicated factors. SEM extends the possibility of relationships among the latent variables and can be viewed as a combination of factor and regression analysis. Real examples where conventional factor analysis is infeasible or inappropriate will be presented as will the results with this data set. The presenters will show how the analyses with the sample data confirmed the hypothesized structure. This data and these findings provide a sound framework for discussing practical implications.

Retesting without a Back-up Form: Implications for Certification Testing

Rachael Tan and Linda Althouse

American Board of Pediatrics, Chapel Hill, United States

e-mail: jinbee@yahoo.com

Developing parallel forms of a certifying examination is challenging, and low candidate volume can threaten the validity of equating. In such instances examinations may be offered infrequently to increase the number of candidates per administration. The examinations in this study are offered every two years. During the most recent administrations, technical issues necessitated same-form retests for affected candidates. The limited research on same-form retest effects within certification testing suggests that repeat examinees are not advantaged by taking the same form of the examination (Raymond, Neustel, and Anderson, 2009). The purpose of this study is to add to this research by examining factors that may have an effect on retest performance and determining whether unwarranted score gains were made due to same-form retesting. Retests were administered to 26 candidates who experienced technical issues during their original examination. When determining whether unwarranted score gains were made by same-form retesting, the following variables were considered: severity of issue, elapsed time between test and retest, and number and direction of changed answers. To control for practice effect, comparisons were made with another examination in which repeat examinees received a parallel form. Item response theory (IRT) was used when possible. On average, candidates changed approximately 25% of their original responses, but less than half were changed from wrong to right, indicating minimal advantage. The severity of technical issues and elapsed time before the retest did not have an effect on the results. Using IRT, in most cases retest scores were within the 95% confidence interval of original scores, revealing no unusual gains. Given the high-stakes nature of certifying examinations, it's crucial that pass/fail decisions are rendered accurately and that results reflect true ability. This research demonstrates that retesting on the same form does not provide an advantage to candidates.

Investigation of Family Background Variables as the Predictors of Mathematic Achievement

Hande Tanberkan and Hayri Eren Suna
Baskent University, Ankara, Turkey
e-mail: handetanberkan@gmail.com

Various variables which are related with students' parents can be effective on the students' achievement. Recent studies show that, diverse family characteristics can be determinative in evaluating the achievement of the students (Hanushek, Kain, Markamn, & Rivkin, 2003; Hess, Holloway, Dickson, & Price). The purpose of the present study is to determine which family-related background variables, which are included in TIMSS background questionnaires, are the significant predictors of Mathematical achievement in 8th grade. All participants which are sampled in Turkey for TIMSS 2011 are used in the study. The family-related variables are gathered form the student questionnaire, school questionnaire and teacher questionnaire together; Mathematical achievement is considered as the means of the plausible achievement values. Multiple regression analysis is used to determine which family-related variables are the significant predictors of the Mathematical achievement in 8th grade. Results will be presented and discussed under the light of broader results of the previous studies on effects of family-related variables on achievement, considering the factors in the Turkish educational system and cultural framework.

Cross-Cultural Development of a Personality Tool for International Use

Louisa Tate, Sarah Mortenson, Katy Welsh and Melanie Brutsche
Cubiks, Guildford, United Kingdom
e-mail: louisa.tate@cubiks.com

The Personality and Preference Inventory (PAPI), currently used in 29 languages, is in the process of being revised in line with the latest psychological research and feedback from experienced PAPI users. The personality scales are grouped into higher-order factors to facilitate understanding and feedback. With the latest revisions, including the addition of two new scales, the aim of this study was to identify an appropriate international factor structure for PAPI3, the revised version of the tool. A cross-cultural approach was taken to confirm the concept for each scale, and an international item-writing team spanning 10 countries was utilised to create new items. For the initial trial, data were collected from a sample of 1,186 individuals across 5 languages. The overall sample was split into two groups, with Exploratory Factor Analysis (EFA) conducted on the first combined language sample. A model for Confirmatory Factor Analysis (CFA) was built on this sample and confirmed on the second sample. The EFA showed a seven-factor solution to be the most robust for group 1, and also the best fit conceptually: Organisation & Structure, Ideas & Change, Interaction, Composure, Impact & Drive, Engagement and Work Momentum. This was confirmed using CFA and acceptable levels of fit for a number of statistics were achieved (median Chi-squared/d.f. = 1.077, CFI = 1, RMSEA = 0.011). Further analyses of the individual language samples were also conducted. A subsequent validation study was conducted on a combined language sample (n = 929), with a Multi-Rater Assessment as the criterion. The range of the predicted moderate / strong correlations was 0.046 to 0.612, with a median correlation of 0.396. While this cross-cultural development approach posed various challenges, the results provide promising evidence that the proposed factor structure is applicable across different cultures and the tool is suitable for international use.

30-Second Interval Performance on the Coding and the Symbol Search Subtests of the WISC-IV: Still WISC Folklore

Lecerf Thierry, Kieng Sotta and Geistlich Sophie

FPSE, UNIGE, Geneva, Switzerland

e-mail: thierry.lecerf@unige.ch

In 1998, Dumont et al. published a paper on the 30-second interval performance on the coding subtest of the WISC-III. This process approach procedure was based on the assumption that for most of children, the number of symbols processed increase over time. These authors found that a decline in the coding score was the norm, and that coding speed was not related with FSIQ. The objective of this study was to apply the 30-second interval recording process approach procedure to the Coding and Symbol Search subtests of the Wechsler intelligence scale for Children – Fourth Edition (WISC-IV). The WISC-IV was administered to 161 non-clinical children aged from 8 to 12 years. For Coding and Symbol Search, the number of items completed were computed for the fourth 30-second interval. Intraindividual means and standard deviation were also computed. Repeated measures analysis of variance were conducted. For Coding, the number of symbols decline from the first 30-second interval to the fourth 30-s interval (11.16, 9.54, 8.75, and 8.35). Coefficients of correlation between FSIQ and Coding scores were weak but all significant, except for the correlation between FSIQ and intraindividual standard deviation. For Symbol Search, the number of symbols decline from the first to the second 30-second interval; then performance remains stable (7.48, 5.43. 5.74, and 5.39). Coefficients of correlation between FSIQ and Symbol search scores were mostly non significant, except for the correlation between FSIQ and the first- and second 30-second intervals. Results indicated that decline was not similar in Coding and Symbol Search, although in both tasks performance was higher in the first 30-second interval. Correlations between FSIQ and intraindividual standard deviation was not significant. We will present the percentage of children who demonstrated "no change, decline, or improve" during each 30-second interval.

Factorial Invariance of Adolescents across Socioeconomic Status (SES) Groups: a Multigroup Confirmatory Factor Analysis

Toni Toharudin[1] and Kai Welzen [2]

[1]Statistics Departmen, University of Padjadjaran, Jatinangor-Indonesia, Indonesia; [2]Radboud University of Nijmegen, Nijmegen, Netherlands

e-mail: toni.t0104@gmail.com

This paper examines how the conditions of high, medium and low SES infuence the behavior of children in answering the 67 items of the Bandung Family Relations Test (BFGT). The 67 items are an operationalization of 6 dimensions derived from Ivan Boszormenyi-Nagy (Boszormenyi-Nagy & Krasner, 1986; Boszormenyi-Nagy & Spark, 1973; Boszormenyi-Nagy & Ulrich, 1981) and Stierlin (1976). The sampling design used in this study was a stratified cluster sample. The sample consisted of 349 pupils of the age 9-12 in Bandung-Indonesia. In order to find out whether the model for three dimensions (Affective Binding, Vulnerability and Justice) was invariant across SES, multigroup confirmatory factor analyses were conducted. Equality constraints were imposed on factor loadings, variances and covariances. Chi-square difference tests indicate that constraining regression weights and factor variances and covariances are equal across SES and leads to statistically significant increases of the chi-square value. Looking at the results, we conclude that Affective Binding is noninvariant across SES for mother and invariant across SES for father. Vulnerability and Justice are invariant across SES for mother and noninvariant across SES for father.Then we are trying to find out in which pair of SES subgroups

the significant differences exists. After that we discuss these differences by looking to the characteristics of the items.

Adjusting a Test to the Population: An Analyis from PAEBES-Alfa's Writing Assessment

Josiane Toledo Ferreira Silva
PUC-Rio and CAEd-UFJF, Juiz de Fora - Minas Gerais, Brazil
e-mail: jositoledo@caed.ufjf.br

In Brazil, the writing assessment, in literacy, is something new, considering the elaboration of a scale using politomic items through the IRT (Item Response Theory). The Avaliação Diagnóstica da Alfabetização do Espírito Santo (PAEBES-Alfa), object of this paper and implemented in 2008, was the first to make this kind of assessment. By examining the results of this assessment, we see that a group of students presented an unexpected performance: a falling proficiency at the end of the second grade. In view of that, we assume that, at the end of the second year, the assessed tasks would have been less complex than those in the tests taken at the end of the first grade. Therefore, this paper is aimed at studying the data produced by the IRT, making analyses concerning adjustments to the population, difficulty and assessed abilities. From these analyses, we were able to built test adjustment curves for each category of response pointed by the correction key, from the most to the least correct. To build test adjustment curves, student percentage by performance level and item percentage were calculated. Each item was placed in level through its anchoring point (which corresponds to the proficiency a student must present to have 65% of chance of getting the item right). The proposed analysis lies on the possibility of having more precise data about test's difficulty as well as detailed information on writing learning processes, specially if we consider that, in Brazil, this kind of assessment will be used in ANA - Avaliação Nacional da Alfabetização, the national literacy test for all third grade's students in the country.

Early Detection of Verbal Comprehension Deficits Using of the Bayley Scales of Infant and Toddler Development, Third Edition

Montserrat Torras Mañá[1], Montserrat Guillamón Valenzuela[1], Ariadna Ramírez Mallafré[1], Carme Brun-Gasca[2] and Albert Fornieles-Deu[2]
[1]Hospital de Sabadell, Sabadell, Spain; [2]Universitat Autònoma de Barcelona, Bellaterra, Spain
e-mail: torras1@gmail.com

The differential diagnosis of specific language impairments (SLI) in ages below 5 is particularly complex because of the close relationship between language development and other cross-sectional skills related to cognition and communication. The importance of early detection of language disorders is justified by the importance of initiating specific therapeutic intervention given the advantage of greater brain plasticity in children in their early ages. The aim of this work is to test the usefulness of the Cognitive and Language scales of the Bayley-III in the early detection of verbal comprehension deficits in SLI. A clinical sample of 187 children with SLI diagnostic hypothesis at 4.5 years was assessed with the Bayley-III before 3.5 years of age and afterwards with other assessment scales of different psychological and psycholinguistic functions in a longitudinal study (MSCA, K-ABC, ITPA). The results indicate that children with SLI scored significantly lower than their control groups in all subtests and compounds of the Bayley-III. Additionally, low scores on the Language composite in the Bayley-III before 3.5 years predicted lower scores in the auditory-vocal channel of ITPA at 4.5 years (AR: R^2 =.364, AA: R^2 =.344, GC: R^2 =.313). A positive correlation was found between the Receptive Communication subtest of the Bayley-III and the Auditory Reception subtest of ITPA (r=.598). A significant correlation was

obtained between the Cognitive Scale from the Bayley-III and the General Cognitive Scale of MSCA (r =.581) and of the Mental Processing Composite of K-ABC (r=.732). We can draw the conclusion that the Cognitive and Language scales of the Bayley-III are a useful instrument in the early detection of verbal comprehension deficits in SLI.

Practice Trials Increases Clinical Utility of a Necker Cube Drawing when Looking for Serious Nonverbal Cognitive Problems

Marberger Tove Kanestrøm and Sundseth Øyvind Østberg

Municipality of Oslo, Oslo, Norway;

e-mail: tove.k.marberger@bsa.oslo.kommune.no

Necker cube drawing is a well-known clinical procedure mostly used as a one trial pass/fail procedure. Despite its relatively low sensitivity for mild cognitive problems in many patientpopulations, the one trial pass or fail probably identify too many with possible cognitive problems in low functioning client groups. In this study a more conservative failure criteria was introduced to check whether the procedure retained useful as a low sensitive, but specific screening. To reduce its sensitivity to "milder" forms of problems we allowed clients up to 5 practice trials to reach a final (sufficient) score. An adult client group (N=215) with various symptoms and ethnical background was assesses with Necker cube drawing and performance measures from WAIS-III. The Necker drawing was scored as near perfect (1 point), sufficient (2 points), not sufficient/loss of 3D impression (3 points) or unrecognizable (4 points). Achieving 1 or 2 points was considered sufficient drawing performance, and 3 or 4 not sufficient/failure. Clients failing their first trial were allowed up to 5 practice trials. The clients were divides into three groups based on their final Necker score. 1. Acceptable after first trial, 2. Acceptable after training, 3. Persistent failure. A WAIS-III POI was estimated (Block design, Symbol Digits and Matrices). 84% (N=70) of the clients with persistent failure obtained an estimated POI below 85, but also in the other 2 groups we found high percentage of est POI below 85 (After practice; 46%/N=23. After first trial; 33.7%/N=28). On the other side when POI fell below 70 (N=36), 27 clients showed persistent failure. Allowing up to 5 trials increased the clinical utility of a pass or fail procedure on a Necker cube drawing when aiming to look only for the most serious nonverbal cognitive problems in a low functioning client group.

A Psychometric Analysis of the Quick Inventory of Depressive Symptomatology-Self Report (QIDS-SR16) in Spanish Patients

Joan Trujols[1], Javier De Diego-Adeliño[1], Albert Feliu-Soler[1], Ioseba Iraurgi Castillo[2], Dolors Puigdemont[1], Enric Alvarez[1], Víctor Pérez[3] and Maria J Portella[1]

[1]Hospital de la Santa Creu i Sant Pau, Barcelona, Spain; [2]University of Deusto, Bilbao, Spain; [3]Parc de Salut Mar, Barcelona, Spain

e-mail: jtrujols@santpau.cat

Psychometrically sound and time-efficient scales that measure depressive symptoms are essential for research and clinical practice. This study was aimed at exploring psychometric properties of the Spanish version of the Quick Inventory of Depressive Symptomatology-Self Report (QIDS-SR16) in a clinical sample. Participants were 173 patients (65% women), mostly with a mood disorder as their primary diagnosis (69% diagnosed as major depressive disorder). Depressive symptoms were assessed by means of the QIDS-SR16 and two interviewer-rated instruments: the 17-item Hamilton Depression Rating Scale (HDRS17) and the Clinical Global Impression-Severity (CGI-S) scale. In addition, three self-report instruments tapping into quality of life, happiness and social support were administered: EQ-5D visual analogue scale

(EQ-5D VAS), Subjective Happiness Scale (SHS), and Multidimensional Scale of Perceived Social Support (MSPSS). Dimensionality, internal consistency, construct validity, criterion validity, and responsiveness to change of the QIDS-SR16 were explored. Exploratory and confirmatory factor analyses replicated the original one-factor structure. The Spanish version of the QIDS-SR16 exhibited good to excellent results for internal consistency (Cronbach's alpha=0.88), for convergent validity [HDRS17 (r=0.77), CGI-S (r=0.78)], and for divergent validity [EQ-5D VAS (r=–0.78), SHS (r=–0.72), MSPSS (r=–0.38)]. The ability of the QIDS-SR16 to identify patients in remission was high (area under ROC curve=0.93). The QIDS-SR16 cut-off point that maximized the sum of sensitivity and specificity in identifying remission was .9. At this cut-off point, sensitivity of the QIDS-SR16 was 80.37% and specificity, 89.09%. Responsiveness to change was also highly satisfactory: patients with greater clinical improvement, as measured by change in CGI-S, showed a greater decrease in QIDS-SR16 scores ($p<0.001$). The findings suggest that the Spanish version of the QIDS-SR16 is valuable as a brief and psychometrically sound self-report instrument to assess depressive symptoms in research and clinical practice.

Psychometric Properties of Travelers Needs Questionnaire

Tatjana Turilova-Miscenko[1], Jelena Levina[2] and Jelena Kolesnikova[3]

[1]University of Latvia, Riga, Latvia; [2]International Higher School of Practical Psychology, Russia; [3]Riga Stradins University, Riga, Latvia

e-mail: tatjana.turilova@inbox.lv

The purpose of this research was to develop the Russian version of Travelers Needs Questionnaire (TNQ) that measures travelers' needs and to determine its psychometric properties. The TNQ was developed for travelers with native Russian language from different countries: Russia, the Baltic states, Belarus, Ukraine. The sample consisted of 160 participants aged from 18 to 68 years (M = 31.17, SD = 9.29) (male – 44%, female – 56%). The factorial validity of the TNQ was established using principal components analysis with varimax rotation; this yielded seven factors: Personal Development (k = 9), Physical Hedonism (k = 9), Pilgrimage (k = 6), Cultural Development (k = 5), Professional Realization (k = 4), Communication and Social Recognition (k = 7), Sport (k = 3). All the TNQ scales had high internal consistency (Cronbach's alpha varied from .76 to .90). The reaction and discrimination indices satisfied the accepted psychometric criteria. The psychometric properties of TNQ satisfied the criteria. The further stage of TNQ development would be the confirmatory factor analysis in broader international sample, the concurrent and convergent validity establishing, and test-retest reliability examination.

Development and Validation of a Measure for Assessing Personal Initiative in Educational Field

Imanol Ulacia, Nekane Balluerka and Arantxa Gorostiaga

UPV/EHU, Donostia, Spain

e-mail: imanol22@hotmail.com

Personal initiative characterizes people who are proactive, persistent and self-starting when facing the difficulties that arise in achieving goals. Despite its importance in the educational field there is a scarcity of measures to assess students' personal initiative. Thus, the aim of the present study was to develop a questionnaire, in Basque language, to assess personal initiative in the academic environment and to validate it for adolescents and young adults. The sample comprised 278 vocational training students (194 men and 84 women; Mean Age = 19.88, SD =

2.93). The questionnaire showed a factor structure including three factors (Proactivity-Prosocial behaviour, Persistence and Self-Starting) with acceptable indices of internal consistency. The questionnaire showed some evidence of convergent validity with respect to the Self-Reported Initiative scale. Evidence of external validity was also obtained based on the relationships between personal initiative and variables such as self-efficacy, enterprising attitude, responsibility and control aspirations, conscientiousness, emotional intelligence, and academic achievement. The results indicate that this new measure is very useful for assessing personal initiative among vocational training students.

Brazilian Confirmatory Factor Analysis of the Utrecht Work Engagement Scale

Felipe Valentini and Maria Cristina Ferreira

Salgado de Oliveira University (Universo), Niteroi, Rio de Janeiro, Brazil

e-mail: valentini.felipe@gmail.com

Work engagement refers to the positive and motivational personal felling in the workplace. This variable is a new trend in occupational psychology, and highlights the well-being rather than the pathological aspects of the work environment. Theoretically, engagement consists of three dimensions: Vigor, Dedication and Absorption. Vigor represents the high level of energy; Dedication is related to the work involvement and the feelings of enthusiasm; Absorption is associated to a full concentration state in the work context. However, the model is recent and needs further empirical evidences. Furthermore, some researchers found evidences of only one dimension and others proposed a model with high correlations among the factors. The present study aims to contribute to this discussion proposing a bi-factor model, setting one general factor as well as three specific factors. For this propose, we used three data base from different researches in Brazil. These researches applied the Utrecht Work Engagement Scale (UWES – 17 items version), and assessed 945 subjects in total. We performed a confirmatory factor analysis based on polychoric correlation and WLSMV (Weighted Least Squares Mean and Variance Adjusted) estimation method. We compared the bi-factor model with the one factor model, and the three correlated factors model. The one factor and the three factors models fitted the data and showed quite similar fit indexes (TLI = .94; RMSEA = .10). Further, the correlations among the three factors were higher than .95. The bi-factor model slightly improved the goodness of fit (TLI = .95; RMSEA = .09). Moreover, all items loaded on the general factor, and 78% of the items loaded on the specific factor as well. These results support a general dimension, and also highlight the importance of the specific factors. In conclusion, the model to comprehend the work engagement might be more complex than the theoretical proposed.

Estimation of Item Parameters in Different Sample Sizes

Felipe Valentini[1], Nelson Hauck Filho[2] and Josemberg Moura De Andrade[3]

[1]Salgado de Oliveira University (Universo), Niteroi, Rio de Janeiro, Brazil; [2]São Francisco University (USF), Itatiba, Brazil; [3]Paraiba Federal University (UFPB), João Pessoa, Brazil

e-mail: valentini.felipe@gmail.com

Item Response Theory (IRT) provides researchers with the possibility of using categorical items in order to estimate parameters of continuous latent variable models. However, small sample sizes tend to reduce the stability of the IRT estimates, and controversies surround the issue of what would be the necessary and sufficient minimum sample size to obtain stable item parameters estimates. Ordinarily, good items (highly discriminative) show good fit even when

estimating their parameters with small samples. In this context, the present research aims to assess the stability of estimates of a (discrimination), b (difficulty) and c (guessing) item parameters by using real data with different sample sizes. We employed a database (12000 subjects) from a Brazilian large-scale educational assessment, comprising14 multiple-choice items about general educational practices. Using bootstrap, we generated five new datasets for each of five distinct sample sizes: 50, 100, 200, 500, 1000 and 5000 subjects. We then averaged parameter estimates for datasets with the same sample size, in order to compare these estimates with those yielded for the full sample. We modeled data using the three-parameter logistic model (3PL) with marginal maximum likelihood (MML) method. Results revealed differences above .10 between discrimination parameter estimates of paired sample sizes below 500 subjects. The difficulty parameter was the most influenced by the sample size. Datasets of 200 subjects (or below) yielded difficulty estimates diverging by around .60 from the full sample estimates; these differences decreased to .30 for the sample size of 500 subjects. Regarding the guessing parameter, all differences were below .10. In conclusion, findings suggest that sample sizes of at least 500 subjects might be enough to estimate a, b and c IRT parameters.

Evidence of the Construct Validity of the Abstract and Spatial Reasoning Test

Felipe Valentini[1], Jacob Arie Laros[2], Renata Manuelly Feitosa De Lima[2], Ronnielison Loiola De Jesus Tavares[3], Wladimir Rodrigues Da Fonseca[3] and Laizza Silva Morais[3]

[1]Salgado de Oliveira University (Universo), Niteroi, Rio de Janeiro, Brazil; [2]Brasilia University, Brasília, BR, Brazil; [3]IESB, Brasília, BR, Brazil

e-mail: valentini.felipe@gmail.com

Fluid intelligence refers to the ability to solve new problems without previous knowledge. This psychological construct is related to the ability to learn and predicts many other important variables such as academic and work achievement. In the present study the construction of a short measure of fluid intelligence in Brazil is described and evidence of its construct validity is presented. The Abstract and Spatial Reasoning Test (ASRT) consists of 24 items distributed over two latent constructs: abstract reasoning and spatial reasoning. Items parameters were estimated by the 3-PL IRT model, and the structure of the instrument was assessed by confirmatory factor analysis. The sample consisted of 1,069 students varying in age from 11 to 17 years. Pursuing the convergence of the model, we constrained the pseudo-guessing parameter (c) at .20 (for items with 5 alternatives) and at .25 (for items with 4 alternatives). The mean discrimination parameter (a) showed an approximate value of 1.00 (a varying from .43 to 1.63), and the mean difficulty (b) was equal to .70 (b varying from -.67 to 1.87). The test information curve suggests that the instrument provides more reliable scores for teenagers with thetas between 1 and 2. Confirmatory factor analysis indicated that the structure of the instrument allows the estimation of two first-order factors (abstract reasoning and spatial reasoning) as well as a second-order factor related to fluid intelligence (TLI=.96; CFI=.96; RMSEA=.020). Concluding, the results indicate that the instrument is appropriate for professionals and in particular for researchers.

Sex Differences in General and Broad Cognitive Abilities for Children

Decaluwé Veerle[1], Tierens Marlies[1], Annemie Bos[1] and Magez Walter[2]

[1]Thomas More University College - Applied Psychology, Antwerp, Belgium; [2]Coordination team Antwerp for Psychodiagnostics (CAP vzw), Brassschaat, Belgium

e-mail: veerle.decaluwe@thomasmore.be

For over a century, sex differences in intelligence have been a topic of great interest among researchers. However, empirical findings are inconsistent and rather confusing. The aim of this study is to investigate sex differences among a Flemish population of children and adolescents. This research focuses on two primary questions: 1) Are there statistically significant and meaningful sex differences in mean levels of the broad cognitive abilities measured in this research? 2) Is there a statistically significant and meaningful sex difference in mean levels of a higher-order (g) factor? Sex differences in general and broad cognitive abilities will be investigated for children by means of a Dutch Cattell-Horn-Carroll (CHC) based Cognitive Ability Test (CoVaT-CHC). The CoVaT -CHC is a new CHC-based intelligence battery for Flemish children and adolescents (Dutch – speaking Belgians). The test measures five broad cognitive abilities (fluid intelligence, crystallized intelligence, short-term memory, visual processing and processing speed). The CoVaT-CHC is designed to measure specific individual cognitive strengths and weaknesses as well as general intelligence. The sample consists of approximately 2000 children, ages 10 to 14. Sex differences will be discussed in the context of previous findings.

Relations between Stress, Professional Maturity and Quality of Life in High School Brazilian Students

Luiz Ricardo Vieira Gonzaga, Claudiane Aparecida Guimarães, Andressa Melina Becker Da Silva, Micheli Aparecida Gomes Dos Santos and Sônia Regina Fiorim Enumo

PUC - Campinas, Uberlândia - MG, Brazil

e-mail: claudianeguimaraes@yahoo.com.br

The transition to College requires adjustments and maturity, but can be stressful for students and compromise their quality of life. This study investigated the level of stress and quality of life, beyond the maturity to career choice for young high school students, analyzing their relationships. Twenty students participated (14 girls), between 16-18 years old, enrolled in the 12th grade, in a public school in São Paulo, Brazil. They answered collectively three instruments: the Lipp's Inventory of Stress Symptoms for Adults (ISSA), which assesses levels, type (psychological, physiological and psychophysiological) and stage of stress - Alert, Resistance, Almost Exhaustion, Exhaustion; the Maturity Scale for Professional Choice (MSPC), and the Quality of Life Inventory (QLI), which evaluates the quality of life of the individual in social, emotional, health and professional areas. The ISSA showed 35% of students at the Endurance phase of stress, with psychological symptoms (30%). An "average" level of maturity for career choice (40%) was identified, but 20% of students in the "superior" level in the MSPC. In the QLI, there were fewer indicators of quality of life in the areas: health (65%), professional (55%), affective (20%) and social (15%). The Spearman correlation showed a significant inverse relationship between overall quality of life and stress symptoms ($p = 0.036$). By Chi-square test, we found significance only between age and overall quality of life ($p = 0.035$), and students with 18 years old presented the best indicators of quality of life. Especially for educational institutions, these data suggest that is important to know how stress acts on the student, which symptoms - physical or psychological - are most prevalent, and which external stressors influence in this context, to aide the career choice by the students.

Aculturation and Personality in Chilean Mapuche Adolescents

Eugenia V. Vinet, José L. Saiz and Natalia Salinas-Oñate
Universidad de La Frontera, Temuco, Chile
e-mail: eugenia.vinet@ufrontera.cl

The Millon Adolescent Clinical Inventory (MACI; Millon, 1993) is a test developed in United States that measures three aspects of individual psychological functioning: Personality Patterns, Expressed Concerns related to adolescence development, Clinical Syndromes with high prevalence in adolescents. There is a MACI Chilean version with adequate indexes of reliability and validity and specific norms developed for Chilean adolescents. Since indigenous groups experience acculturative changes resulting from contact with the non-indigenous majority and MACI's Chilean version has not been tested on these groups, this study examined the relationship between two aspects of acculturation (involvement in indigenous culture and involvement in the non-indigenous mainstream culture) and the psychological areas measured by the MACI. A sample of 211 Mapuche secondary students (96 men) from La Araucanía, a Chilean geographic region which has a high rate of indigenous population, answered Mapuche Acculturation Scale (EAM) and MACI's Chilean version. A series of canonical correlations showed interpretable multivariate results, but only in the female sample. Association between EAM and Personality Pattern scales indicates that when there is a greater participation in the Mapuche culture and lesser participation in the non-indigenous culture, a pattern of personality functioning characterized by inhibition and submission is presented. Moreover, when considering the MACI's Clinical Syndromes scales, the same EAM pattern relates to high scores in disorders characterized by dysphoric feelings and lower scores in disorders characterized by externalizing behaviors such as substance abuse, impulsivity and delinquency. No interpretable results emerged in men sample and when Expressed Concerns scales were considered. Results are understood in terms Mapuche culture; the MACI Chilean version is evaluated as a suitable tool for personality assessment in Mapuche adolescents if cultural differences from the mainstream Chilean culture are considered when interpreting test results.

Comparing the Rating Effectiveness of Personalized vs. Non-Personalized Feedback to the On-Line Raters of English Speaking and Writing Assessment

Alex Volkov[1], Kristina Chang[1], Jake E. Stone[1], Michelle Y. Chen[2] and Amery D. Wu[2]
[1]Paragon Testing Enterprises, Vancouver, Canada; [2] The University of British Columbia, Vancouver, Canada
e-mail: volkov@paragontesting.ca

This study investigates the effects of type of feedback to the raters (personalized vs. non-personalized, and control) on their performance in rating the constructed responses of the speaking and writing assessment. Even though there is plenty research on initial rater training (e.g., Wang, 2010), methods for on-going feedback and calibration are not sufficiently studied. Personalized rater performance reports are costly and have yielded mixed results as to their effectiveness (Elder, Knoch, Barkhuizen, & von Randow, 2005). The speaking and writing components of the Canadian English Language Proficiency Index Program- General (CELPIP-G) are part of a large scale, standardized assessment for high-stakes immigration and citizenship purposes. Test takers construct their own written and verbal responses to the task requirements. Responses to each task are rated by at least two independent raters on a scale of 1-5 on four dimensions of proficiency. Rating assignments are managed through a centralized online rating system. Many-facets Rasch measurement (MFRM) model was used to identify underperforming raters. The identified raters were randomly assigned to two feedback-receiving methods. With the personalized method, ratings on all four dimensions by a specific underperforming rater,

along with the original response, are juxtaposed with those of the same response evaluated by the benchmark (calibration) raters who have shown strong validity and reliability in rating. With the non-personalized method, only exemplar ratings from benchmark raters are shown to the underperforming raters. Remaining raters function as the control group. Ratings under different experimental conditions will be recorded weekly and feedback provided until the time when the underperforming rater has met pre-specified calibration criteria or until the end of February. The effect of the feedback method will be evaluated by the times of feedback until calibrated and the weekly changes in performance (MFRM analyses and exact and adjacent agreement).

Factor Structure Analyses of the Center for Epidemiologic Studies Depression Scale: Applying Bayesian Structural Equation Modeling Approach

Li Wang, Yulan Qing, Richu Wang, Chengqi Cao and Jianxin Zhang

Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

e-mail: wangli1@psych.ac.cn

The original four-factor structure of the Center for Epidemiologic Studies Depression Scale (CES-D) has been argued since formulated in 1977 resulting in several alternative structure models. Several authors suggested that the lack of consensus in the structural model was due to the limitation of traditional methodology, such as maximum likelihood Confirmative Factor Analysis (ML-CFA). With a strict pre-assumption of independent (zero correlation) between factors, ML-CFA can lead to poor model fits and distorted factor structures with overly restrictive constraints on exact zero cross-loading and residual correlation. Bayesian structural equation modeling (BSEM) addressed this limitation by allowing cross-loading and small residual correlations between factors. The current study investigated the factor structure of CES-D using BSEM approach. Four alternative models were investigated applying BSEM among a sample (N=326) of Chinese survivors who lost their children in Wenchuan Earthquake. Result: Each of the 4 alternative models only fitted well when approximate zero cross-loadings and residual correlation were specified at the same time. The original four-factor model with approximate zero cross-loadings and residual correlation was preferred. BSEM approach helped with further insight on the latent structure of CES-D and the on-ignorable interrelations between depressive symptoms.

Parameter Equivalent Comparative Study of Testlet-Based Tests Using 3PLM and 3PTM

Yi Wang, Min-Qiang Zhang and Wen-Qing Tang

Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou, China

e-mail: wangyipsychology@163.com

Along with the development of research on examination, testlets appeared in different examinations increasingly. When item response theory (IRT) models are applied in test equating, strong statistical assumptions-local independence (LI) must be met. However, previous studies have shown that local independence is likely to be violated when testlets are contained in test. Hence, when equating tests composed of testlets, that local dependence is ignored can lead to distortion of equating coefficients using standard IRT model. In order to solve this problem, this paper use a testlets-based model-3 Parameters Testlet Model (3PTM), which is derived from IRT 3 Parameters Logistic Model by adding random-effect parameters associated with each testlet. The findings of equating tests made up of testlets using 3PTM were compared with standard IRT model-3PLM, which does not account for local dependence among items from a common testlet. IRT characteristic curve method and specific procedure for calculating equating

coefficients are based on Wilcoxon sign-rank test, a lot of experiments are done using Monte Carlo simulation method. The effectiveness of equating tests containing testlets was investigated in terms of the accuracy of estimation of item parameters (AEIP), the number of examinees and the degree of local dependence. Results suggest that 3PTM is better than 3PLM in recovery and has significant differences mostly, so 3PTM is suitable for equating tests based testlets. In addition, we know that the C parameter can affect participants' answer as same as random-effect parameters. We use 3PTM to delve into it and find that error change without regularity with the increase of C parameter. But the equivalent effect is better than 3PLM. This paper provides that 3PTM performs better than 3PLM when equating tests composed of testlets. For other influence of the factors may need further research.

Factor Structure and Validity of Future-Oriented Coping Inventory Among Chinese University Students

Yu Wang, Yiqun Gan and Xiangrong Yang
Peking University, Beijing, China
e-mail: wy13131@126.com

In a world full of changes and crises, young people are facing ever-increasing challenge and change in their lives. In this context, future-oriented coping is essential to psychological adaption and stress management. The existing related measures for future-oriented coping mainly focuse on aging problem in middle-aged and old people, or assess the some stages or aspects of proactive coping competence. This research aimed to design a Future-Oriented Coping Inventory for Youth (FCIY)by integrating motivation, cognitive competence and action processing into the construction of future-oriented coping. Based on result of the Exploratory Factor Analysis in 185 university students, the FCIY presented four dimensions: future-orientation, insight ability, resource accumulation, and planning. Reliability and validity were tested in another sample (N=211),. Rresults indicated high internal consistency (alpha coefficient of the total scale and subscales were 0.73~0.87). Confirmatory Factor Analyses demonstrated good fit of the four-factor model (χ^2/df =1.96, CFI=0.98, NNFI=0.98, RMSEA=0.064). FCIY was significantly correlated with Proactive Coping Inventory (proactive coping dimension: r=0.59, p<0.01; preventive coping dimension: r=0.57, p<0.01). To obtain criterion validity, FCIY was administered among graduates looking for jobs (N=189). Structural Equation Modeling indicated that the four dimensions of FCIY represented the four sequential stages of future-oriented coping, resulted in lower the job-seeking anxiety (χ^2/df =1.62, CFI=0.98, NNFI=0.97, RMSEA=0.057). Findings suggest that the FCIY possesses good psychometric propertied, and could be used to measure future-oriented coping in youth population.

Equating Challenges when Switching to a New College Admission Test

Jonathan Wedman and Marie Wiberg
Umeå University, Umeå, Sweden
e-mail: jonathan.wedman@edusci.umu.se

Equating a college admission test that should be valid for several years is always a demanding task because of test security issues and potential changes in the group of test takers. When such a test is altered by adding several new subtests and items, and by changing the scoring scale, then equating is especially challenging because the test score needs to be fair both for those who take the new test and those who took the old test. There are several equating and linking designs available but no one is indisputably superior in this particular situation. The aim of this study was to examine how a previously used college admission test, which remains valid for five years after

switching to the new test, should be equated and linked to a new college admission test (with added items and subtests) in order to keep the validity of the test scores for all test takers who will apply for university. The hypothesis was that the different equating designs and methods would yield similar test taker rankings. The equating designs included equivalent groups, nonequivalent groups with anchor test and the current practice of the college admission test. Also, different observed score equating methods were used. Apart from the data from the old and the new tests, simulations were conducted to examine how the designs and methods work under different conditions. Preliminary results from analyzing test taker data indicated that test takers would obtain slightly different test scores but similar scale scores over most designs and methods. Preliminary results from simulated data indicated a notable variation in test scores and scale scores depending on choice of equating design and method under certain conditions. The results will be discussed in a larger perspective including limitations and suggestions for future research.

Examining the Structural Equivalence of English and German versions of the TestAS

Frank Weiss-Motz

TestDaF-Institut, Bochum, Germany

e-mail: frank@weiss-motz.de

The TestAS (Test for Academic Studies) is a newly developed standardized test measuring intellectual abilities important for entering institutions of higher education in Germany. The test consists of a core test that measures general cognitive abilities and four subject-specific test modules to choose from. Test items vary from completely language-free tasks like "Continuing Numerical Series" to highly language-related ones like "Inferring Relations". The test itself is offered and administered in two languages (English, German). The objective of this study was to compare different methods of examining the difficulty equality as well as the structural equality of the different language versions. In contrast to the equality in difficulty, the structural equality focuses on interrelations of items and scales. Therefore DIF analyses as well as factor analyses with procrustean factor rotation and SEM methods were carried out. Results show that though equality of difficulty is hard to achieve, achieving a structural equality is even more so. Even tasks that are apparently language free can show minor differences in difficulty and factor structure between the language versions. The more language related a task is, the lower the structural equivalence proves to be. More research needs to be carried out on this interaction.

Metaphorical Competence Assessment – A Pilot Study

Katarzyna Wiejak, Grazyna Krasowicz-Kupis, Katarzyna Maria Bogdanowicz and Dorota Kwiatkowska

Educational Research Institute, Warsaw, Poland

e-mail: k.wiejak@ibe.edu.pl

The presented study is a part of IBE Dyslexia Project, which concerns the early identification of specific reading and spelling disorders, specifically the risk of developmental dyslexia. The main goal is to develop a battery of tests to measure reading and spelling abilities in children before starting and at the beginning of formal reading instruction, that are predictive of these disorders. Polish studies have shown that a large group of children with dyslexia experience a different kind of language deficits, and communication problems. Apart from the obvious problems with reading and writing symptoms of dyslexia include semantic and metasemantic deficits (Krasowicz-Kupis, 2008). On the other hand, research in cognitive linguistics shows that metaphor is an essential ingredient of communication and metaphor comprehension is framed as a metasemantic ability (Gombert 1990). So far, little research on the metaphorical competence

of Polish children with dyslexia was conducted. One of the reasons is the lack of tests to measure these abilities in Polish. The aim of this study is to examine an Metaphorical Competence Test. Metaphorical competence was assessed on the basis of a 5 tasks as multiple-choice task and fill in tasks (sentences and short stories) that measures ability to recognize, produce and comprehend figurative utterances. Poster presents the results of pilot study carried out on a group of 400 children aged 6-8 years. Results and additional aspects of validity and reliability will be presented in this study.

Using Signal-Detection Theory to Measure Cue Recognition in Multiple-Response Items

Ada Woo[1], Will Muntean[2] and Joe Betts[2]

[1]National Council of State Boards of Nursing (NCSBN), Chicago, United States; [2]Pearson Vue, Chicago, United States

e-mail: adawoo@outlook.com

Decision-making skills are increasingly recognized as important in many areas of academia, employment, licensure, and certification. This has resulted in a need to assess discrimination between relevant and irrelevant cues in service of problem solution on academic and licensure/certification tests, e.g. those that evaluate the candidate's decision-making ability. Traditional methods of evaluating cue recognition have relied upon multiple choice (MC) items and scoring models based on item response theory (IRT). This research will investigate two potentially productive areas in the assessment of cue recognition for decision-making. First, the advantages of a more flexible item type that allows for multiple-responses (MR) will be investigated. MR items allow a number of responses to a single problem set to be evaluated, requiring candidates to recognize multiple important cues while simultaneously excluding irrelevant ones. Second, a signal-detection theory (SDT) methodology for scoring will be constructed, which can separately estimate cue discrimination (i.e., the ability to select appropriate cues from inappropriate cues) and the propensity to over/under endorse cues (i.e., to select too many or too few cues, regardless of their appropriateness). This makes it distinct from IRT models. A large item pool of MR items field-tested in a sample of professional certification candidates will be used. These items will allow for the evaluation of cue recognition for diagnostic decision-making. Items will be analyzed using traditional item response models along with the proposed SDT model to allow for comparisons. Results may indicate that SDT will provide a plausible model for evaluating responses to MR items that assess cue recognition in a decision-making scenario. The results of this research will provide a potentially flexible—theory driven—approach to assessing cue recognition that can be applied in both research and testing programs.

Understanding Test-Taking Strategies – A Validation Study of the CELPIP-G English Reading Comprehension Test

Amery Wu[1] and Jake Stone[2]

[1]The University of British Columbia, Vancouver, Canada; [2]Paragon Testing, Burnaby, Canada

e-mail: ameryw@yahoo.com

Messick (1995) highlighted the substantive aspect of construct validity, which included "the process models of task performance, along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks" (p. 745). Providing validity evidence as highlighted by Messick, the purpose of this study is to investigate test takers' cognitive engagement when answering the reading comprehension items of the Canadian English Language Proficiency Index Program- General (CELPIP-G) Test. A sample of 189 adults from different cultural/language background participated in a pilot test of the CELPIP-G reading

test consisting of four item types.The multiple-choice questions within each type required the same task (e.g., choosing the correct summary for the paragraph) in response to the same passage. The four item types were developed to have different difficulty levels targeting different levels of test-taker ability. Immediately after completing each item type, test takers reflected on whether they engaged in 10 pre-specified test-taking strategies. Six of the strategies were believed to be construct-relevant cognitive engagement (e.g., understanding main ideas; translating). The remaining four were believed to be construct-irrelevant testwiseness for multiple-choice items (e.g., matched words in the passage and options). It was hypothesized that, to show evidence of construct validity, test takers should engage mainly in construct-relevant cognitive strategies when answering item types that were less difficult than their ability, but would also engage in testwiseness strategies when answering item types that had difficulty beyond their ability. Data were analyzed using logistic regression. The results generally confirmed our hypotheses. The predicted probabilities of engaging in construct-relevant cognitive strategies were generally high (0.5 to 0.9) and were relatively low in construct-irrelevant testwiseness strategies (0.05-0.50). However, the variation in probability of both kinds of strategies was evidently an interaction between the item-type difficulty and test-taker ability.

Comparing the Reliability and Validity of Short-Form Five-Factor Personality Tests

Jianping Xu, Hongyan Li, Jiyue Chen and Yexin Fan

Insititute of Psychological Measurement and Human Resources, School of Psychology, Beijing Normal University, Beijing, China

e-mail: xujp@bnu.edu.cn

The five-factor model (FFM) of personality is the predominant model in trait psychology. The gold standard for measuring the dimensions of the FFM is the 240-item NEO-PI-R. However, this instrument is too lengthy for many research and applied uses. Therefore, several short-form instruments were developed, including NEO-FFI (60-item), 44-item Big Five Inventory (BFI-44), Mini-Markers Set (MMs, 40-item),10-item Big Five Inventory (BFI-10), Big Five Locator Questionnaire (BFLQ, 25-item), Ten Item Personality Inventory (TIPI, 10-item), Single-Item Measures of Personality (SIMP, 5-item). The objectives of this study were to compare the reliability and validity of all these short-form tests in a research, and to find evidences to balancing the demands of brevity versus reliability and validity. The sample was composed by randomly selected 352 students in Beijing, including post graduate student, college student, minority student and overseas student. The subjects finished NEO-FFI, BFLQ, MMs, BFI-44, BFI-10, SIMP, TIPI and a personality criterion at a time. Two weeks later, we did the retest. The different tests in terms of test–retest reliability, factor structure, and correlation with the same personality criterion were evaluated. Together the analyses suggest that the longer tests, for example NEO-FFI, BFI-44 shows higher test-retest reliability and correlation with the criterion; tests with fewer items also meet the psychometric index. Overall, the longer tests have higher correlation with the criterion, but for some single dimension, the shorter tests demonstrated a closely similar pattern of criterion correlations when compared with the longer scales, some are even higher than the longer ones. For confirmatory factor analysis, the shorter scales own better model fitting. The shorter tests with 5 or 10 items offer a reasonable alternative to longer scales when research is not for clinical analysis or personality is not the primary concern.

The Development and Validation of Social Competence Inventory: A Confirmatory Factor Analysis

Gonggu Yan

Beijing Normal University, Beijing, China

e-mail: gregyan@bnu.edu.cn

Social competence is a complex, multidimensional concept consisting of social, emotional cognitive, and behavioral skills, as well as motivational and expectancy sets needed for successful social adaptation. Social competence also reflects having an ability to know oneself and others, learn from past experiences, and apply that learning to the changes in social interaction (Semrud-Clikeman, M., 2007). Despite its popularity and usefulness in the real life, the factorial structure of the Social Competence remains controversial. The present study developed a Chinese version of Social Competence Inventory (SCI) which is a self-reported online questionnaire with 100 items and examined the reliability, factor structure, and validity of SCI scores collected from a sample of 3760 Chinese workers, age from 18 to 60 years old, job title from staff to executive in the organizations. Results in general support that the scores measured from SCI reflected the hypothesized factors and correlated with the position level in the organization. Internal consistencies observed across the different groups studied were high. Implications for assessment and future directions are discussed.

Is the Same Best Practice Applicable to All Cultures? —Lower Reliability Associated with Negatively-Keyed Items when Testing in Chinese

Tanya Yao

CEB- SHL Talent Measurement, Surrey, United Kingdom

e-mail: tanya.yao@shl.com

It is considered best practice to balance positively-keyed and negatively-keyed items in questionnaires. However, little research has been done to study the effect of such practice. Furthermore, cultural differences and response-bias play a role in how people respond to rating scales. In regards to negatively-keyed items, such differences might distort people's real responses and reduce predictive power and reliability of items. The objectives of the study were to 1) examine the reliability of positively-keyed and negatively-keyed items in Simplified Chinese (CHS) and UK English (UKE) data, and 2) explore whether items of the equivalent content have stronger reliability if worded positively or negatively. Native Chinese speakers (N=395) completed a self-report Motivation Questionnaire (MQ) as part of an employment requirement. The questionnaire was adapted from UK English to Simplified Chinese following ITC Guidelines for Test Adaptation. In Experiment One, the reliabilities of positively-keyed and negatively-keyed items were compared against the UKE reference sample. In Experiment Two, the negatively-keyed items were reversed into equivalent positively-keyed items and trialled together with all items in the original questionnaire with further UKE and CHS speaking participants. The reliability was again examined. Results In Experiment One, 16 out of 17 scales showed higher reliability in positively-keyed than negatively-keyed items in the CHS sample but not the UKE sample. The preliminary results of the second experiment suggested larger improvements in reliability for the re-written positively–keyed items when compared to original negatively-keyed items in CHS than in UKE. Reliability was lower on negatively-keyed than positively-keyed items especially in CHS. The effect of response style and personality traits in rating scale questionnaire response bias has been long discussed; culture might be playing a role here resulting in low reliabilities of negatively-keyed items. The results suggest that the general

best practice of balancing positively and negatively-keyed items may require caution in certain cultures.

Determination of Achievement in Secondary School Mathematics Curriculum Applied in Turkey According to PISA Mathematics Literacy Competency Levels

Emine Yavuz and Sibel Ada
Gazi University, Ankara, Turkey
e-mail: minre_32@hotmail.com

PISA finds out the capacity of students post compulsory education to be able to use the knowledge in daily life, to make mathematical derivations, make comment on problems concerning in varying cases and make derivations from what have been learnt to solve the problems. The countries make educational policies based on results of that exam determining the quality of the compulsory education. In this line Turkey has made some changes in its education program based on PISA 2003 results. With such changes the following basic skills have been incorporated in new secondary school mathematics curriculum: 1. Problem Solving 2. Mathematical Process Skills • Communication • Reasoning • Associating 3. Affective Skills 4. Psycho-motor Skills 5. Information and Communication Technologies. This study has been made to find out the competency level of PISA mathematics literacy of achievement established to provide above mentioned skills in secondary school mathematics curriculum in Turkey. After finding the levels, it has been aimed to find out if the students' PISA results are in parallel with the achievements in the curriculum. The document examination has been carried out for this purpose and authors have coded the achievements according to the PISA mathematics literacy competency level by means of content analysis. Thereafter, the accumulations of achievements in the levels have been determined by use of frequency and percentage analysis. As a result, it is seen that the achievements in secondary school mathematics curriculum in Turkey concentrate in initial four grades of PISA mathematics literacy competency levels and the accumulation is seen in second level. According to PISA 2012 Report, the highest number of students is in second level. As it is understood from the fact that the results are not parallel, it is expected that increasing the levels of achievements will increase the quality of compulsory education.

Missing Data Techniques and Application Status in Longitudinal Studies

Sujing Ye
Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou, China
e-mail: 985299350@qq.com

Missing data are not uncommon in longitudinal studies. Different missing data techniques affect the accuracy of the results and validity of statistical inference. To summary missing data techniques in longitudinal studies and explore problems in theory and applied researches and make recommendations accordingly. This paper first summarized missing data techniques based on different mechanisms of missingness, including Expectation Maximization algorithm, Full Information Maximum Likelihood, Multiple Imputation based on the assumption of missing at random; selection models, pattern-mixture models based on the assumption of missing not at random. Second, we conducted a methodological review of missing-data reporting practices for longitudinal data in the field of psychology in China through literature analytic method. We got the initial sample of 275 journal articles, 41 dissertations on longitudinal studies in China Journal Net database. Through the literature review, we found 59(64.10%) contain a certain degree of missing data. The majority (66.10%) used deletion method, and the rest did not report the

method used. This paper provides a sketch of the missing data techniques and application status in longitudinal studies. The validity of these available missing data techniques needs further investigation and the adoption of the practice of reporting on missing data in published research also need to be done.

Robust Variance Estimations in Cognitive Diagnostic Models

Jung Yeon Park and Matthew Johnson
Columbia University, New York, United States
e-mail: jyp2111@tc.columbia.edu

Recent availability of large-scale assessment data of educational achievement makes it possible to compare the cognitive strengths and weaknesses across countries. Cognitive diagnostic modeling (CDM) is one of the methods used to diagnose the cognitive attributes. We have analyzed the 2007 TIMSS Grade 8 Mathematics data using the multi-group DINA (MG-DINA) model. The problem with application of this model to the TIMSS data is that TIMSS assessment uses a complex sampling design to select samples of student within schools and classrooms. The complex sampling design is different from a simple random sampling in which every student has an equal chance of selection and observations from different students are considered to be statistically independent. Thus, variance estimations by ignoring the sampling design would be generally inaccurate. This research aims to establish more robust methods to conduct variance estimations within CDM framework. First, we use Jackknife Repeated Replication (JRR) method. A total of 75 Jackknife replicated weights for each country allow to approximate variances for all parameters used in MG-DINA. Secondly and more importantly, we develop Sandwich estimator (Huber, 1967; White, 1982) in CDM. These estimators are a widely used tool for variance estimation because of its robustness when observations are correlated but the specification of correlation structure is unclear. Additionally, using these estimators are computationally efficient in large scale data. After fitting MG-DINA model to TIMSS data, we calculate confidence intervals by using the sandwich estimators of variance. We use these confidence intervals to compare estimated knowledge skills at country-level across nine countries selected from Asia, Europe, and the United States.

The Influence of Personality Traits on the Variability of Estimates

Yulia Yusupova
South Ural state University, Chelyabinsk, Russia
e-mail: hopeup@gmail.com

The purpose of the article is to identify the specificity of estimate variability under the influence of contextual factors and depending on the personal traits. The variability as essential property of evaluation is necessary for keeping the adequacy of worldly evaluations and the accuracy of professional evaluations. It is theoretically described the psychological content of estimate variability. Variability of estimates – is the common property of all types of estimates to modify the degree of the evaluated properties due to changes in the object of evaluation, evaluation basis, in comparison and in the form of estimate expression. Researches reveal that evaluations are changing under the influence of many factors. The attempt of work classification of these factors is also presented in the article. It was conducted the experimental procedure for studying how estimates changes. It was developed the procedure of defining the variability index for studying the dependence of variability from personality traits. It was defined that presentation of additional information about individual characteristics of assessed person leads to changes in estimates. Also it was defined that the sign valance and format of information have the great

influence on estimates. Moreover variability of estimates depends on a complex of personality traits. This complex determine the high and low levels of estimates variability.

Evaluating Acquiescence and Negative Wording in Self-Report Testing

Cristian Zanon[1], Claudio Simon Hutz[2] and Markus Zenger[3]

[1]University São Francisco, Itatiba, Itatiba, Brazil; [2]Federal University of Rio Grande do Sul, Porto Alegre, Brazil; [3]University of Leipzig, Leipzig, Germany

e-mail: cristianzanon@yahoo.com.br

Acquiescence is an issue on testing. The way persons respond to scales may increase errors and produce an additional artificial factor on the expected internal structure. Negative wording constitutes a common method to evaluate a construct that can reduce acquiescence, but it can also produce an additional factor when the intercorrelations among negative items are higher than the correlations of negative and positive items. This may be the case of the Life Orientation Test-Revised, whose dimensionality has been controversial in the literature. To investigate these possibilities, we compared three models: a)one-factor model, b)two-factor model, and c)random intercept factor model (Maydeu-Olivares & Coffman, 2006). The late model allows intercepts to vary across subjects. Participants were 524 undergraduates. The two-factor model presented the best fit indices (χ^2=38.6, df=8, p < .001, AIC=23, CFI=.97), compared with the one-factor model (χ^2=49.1, df=9, p < .001, AIC=31, CFI=.95), and with the random intercept factor model (χ^2= 39.8, df=8, p < .001, AIC=31, CFI=.96). However, the two-factor model presented high correlation between the latent variables (r=.90) suggesting negative and positive items might artificially produce two factors. Additionally, acquiescence seems to play a role on these data, and model "c" seems the best model.

Dimensionality of the Core Self-Evaluation Scale: Contrasting Common Factor Models with the Random Intercept Item Factor Analysis

Markus Zenger[1], Cristian Zanon[2] and Andreas Hinz[1]

[1]University of Leipzig, Leipzig, Germany; [2]Federal University of Rio Grande do Sul, Porto Alegre, Brazil

e-mail: Markus.Zenger@medizin.uni-leipzig.de

The Core Self-Evaluation Scale (CSES) measures a higher order latent personality trait that includes several facets of the constructs of self-esteem, self-efficacy, locus of control, and neuroticism. The aim of the present study was to test dimensionality of the CSES, using different factor models (with one, two, and four factors) compared to an approach introduced by Maydeu-Olivares & Coffman (2006). This approach is based on the assumption that common linear coefficients in SEM may be too restrictive when evaluating the underlying dimensionality in self-rating questionnaires. The study is based on a representative survey of the German population (N=2,451). The factorial structure of the CSES was tested using CFA. Following Maydeu-Olivares, an additional latent factor corresponding to acquiescence was estimated in this model, with the relaxed constraint of fixed intercepts across all respondents and the intercepts were allowed to vary between individuals. The one-factor model showed the worst fit to the data (CFI=.780, RMSEA=.138, SRMR=.090). The best fit showed the two-factor model (CFI=.946, RMSEA=.069, SRMR=.040) followed by the four-factor model (CFI=.885, RMSEA=.106, SRMR=.074). Additionally, the random intercept item factor analysis for the one-factor model showed adequate fit (CFI=.943, RMSEA=.071, SRMR=.038), and performed comparable to the two-factor model. Regarding the latter model, factor loadings ranged between .55 and .67 (absolute scores, p<.001) for the CSES factor and between .26 and .34 for the random intercept factor. The variance of the random intercept factor was small with 0.08 (SE=.004). Based on the

results of the CFA, the two-factor-model and the one-factor-model with a random intercept fit the data rather well. Therefore, other criteria (e.g. predictive value for external variables) have additionally to be taken into account, when deciding which solution is preferable. Results will be discussed in the light of further psychometric properties and external validity findings.

Application of Generalizability Theory to International Large-Scale Assessments: A Demonstration with the PISA Data

Wen Zhang[1], Michelle y. Chen[2], Eric k. H. Chan[2] and Bruno D. Zumbo[2]

[1]Université Laval, Québec, Canada; [2]The University of British Columbia, Vancouver, Canada

e-mail: zwlisa@gmail.com

Generalizability theory (GT) is increasingly used in educational assessment to investigate sources of error variance -- e.g., rater variance. GT, however, has been found comparatively rarely used to investigate cross-cultural comparability in large-scale international assessments. With an eye towards helping researchers and policymakers understand how to apply GT to cross-cultural assessments, the purpose of this paper was to demonstrate its application and discuss several psychometric issues that arose when investigating the generalizability of scores from the 2009 Programme for International Student Assessment (PISA) science test across a wide range of jurisdictions. G-studies with multi-faceted designs were used to quantify the source of error variance across a range of jurisdictions from those ranking high, middling, and low on the 2009 PISA science test. Three sources of variance, student, jurisdiction and item, were investigated. In the first two G-studies, student was chosen to be the object of measurement. The first G-study was a mixed design in which students were nested in the jurisdictions and then crossed with the items. In the second G-study, we analyzed each level of the fixed jurisdiction facet separately. In the third G-study, jurisdiction was chosen to be the object of measurement in order to analyze whether it is possible to rank jurisdictions reliably based on the chosen items. The results showed that a global analysis including different jurisdictions and separate analyses for each jurisdiction were useful to investigate and interpret the influence of jurisdiction on test score variability. Different objects of measurement in G-studies led to different results and interpretations. The paper demonstrated that results from G-studies can provide useful measurement invariance evidence when establishing the validity for cross-cultural comparability. This paper will also help researchers, educators, and policymakers understand how to apply GT to large-scale cross-cultural assessments and how to interpret the results.

The Revision and Validation of Academic Motivation Scale in China

Bo Zhang[1], Jian Li[2] and Houcan Zhang[2]

[1]School of Psychology, Beijing Normal University, Beijing, China; [2]Beijing Key Laboratory of Applied Experimental Psychology, School of Psychology, Beijing Normal University, Beijing, China

e-mail: zhangbo@mail.bnu.edu.cn

Self-determination theory (SDT) has contributed a lot in our understanding of human motivation. Based on SDT, Vallerand developed Academic Motivation Scale (AMS) to assess students' motivation to learn. It is composed of three intrinsic motivation (IM) subscales, three extrinsic motivation (EM) subscales and an amotivation subscale. AMS has been applied to educational context successfully in western cultures. However, no psychometrically validated version is available in China. The present study was aimed at the revision and validation of AMS in China. The back-translated version was administered to 882 high school students and 419 vocational students. Retest took place among 67 normal high school students two months later. Cronbach's alpha of subscales ranged from 0.75 to 0.86 and test-retest reliability ranged from

0.57~0.81. CFA demonstrated that the seven-factor model fitted well, □²/df=3.78, CFI=0.93, RMSEA=0.05. In order to further test its construct validity, extrinsic and intrinsic motivation subscales from Motivated Strategies for Learning Questionnaire were chosen as motivational counterparts. Results showed that IM subscales had significant positive correlation with their intrinsic counterpart (r=0.52~0.58), EM subscales had significant positive correlation with their extrinsic counterpart (r=0.34~0.46), and amotivation subscale had significant negative correlation with both intrinsic and extrinsic counterparts (r=-0.37, -0.12). Basic Psychological Needs Scale and The Learning Climate Questionnaire were adopted as motivational antecedents. They were found to show significant correlations with IM subscales (r=0.32~0.52), EM subscales (r=0.13~0.42) and the amotivation subscale (r=-0.32~-0.38). School satisfaction, as motivational consequence, was found to had significant correlations with IM subscales (r=0.53~0.66), EM subscales (r=0.10~0.44) and amotivation subscale (r=-0.58). Group comparison revealed that high school students scored significantly higher than vocational students in IM subscales (t=2.95~6.21), significantly lower in one EM subscale (t=-3.81) and amotivation subscale (t=-8.1). The Chinese version of AMS was psychometrically sound and could be applied in China.

Effects of Relationship Type, Thinking Mode and Decision Style on Risk Preference

Yufeng Zhang and Hong Li
Tsinghua University, Beijing, China
e-mail: yufengzhang.tsu@gmail.com

The purpose of the present research is to explore and compare the effects of intuition and deliberation on decision-making risk preference under different conditions of relationship type between rescuers and the trapped in emergency situations. The research consists of two studies, and 68 and 93 participants are recruited for study 1 and study 2 respectively. A 3 (relationship type: intimate, ordinary and strange) ×2 (thinking mode: intuition and deliberation) mixed design is adopted in study 1, and a 3(relationship type: intimate, ordinary and strange) × 2 (thinking mode: intuition and deliberation) ×2 (decision style: intuitive and deliberative) mixed design in study 2. The main findings are: (1) The closer the relationship is, the higher the risk preference is. (2) There are no significant differences in decision-making risk preference between the two modes of thought (intuition versus deliberation). (3) Intuitive participants reveal higher risk preference than deliberative ones. (4) The interaction of thinking mode and relationship type is significant, and it further suggests that relationship type has a significant effect with intuition rather than with deliberation. In conclusion, intuition and deliberation affects decision-making risk preference under different conditions of relationship type in emergency situations.

Social Networking Behaviors, Online Social Capital and Influence of Chinese Culture: A Survey on 450 Chinese Elite Undergraduates

Zheng Zhang
Tsinghua University, Beijing, China
e-mail: zhangzheng@tsinghua.edu.cn

Online communication can help users improve communication skills, social involvement, mental health, and Internet use is positively correlated with social capital. (Ellison et al., 2007) Through researches on Facebook use and social capital of college students, we may find that the more users use Facebook, the better could they make and maintain social connections with friends. In this research, we investigated 450 undergraduates in Tsinghua University and Peking University,

which are be seen as the top two universities in China. Based on the data on their Renren.com (The Chinese version Facebook) using behaviors, the analysis showed that high using intensity means that people use Renren.com to post a lot and communicate with others frequently. These behaviors enlarge the visiting number on their homepage, extending their social relations and improving their social capital. We also gathered data using CPAI-II, showing that Face dimension and Defensiveness dimension both could indicate the high social capital. Self-disclosure online could improve trust and shorten psychological distance, which can also improve the online social capital of college students. Also we may discuss what role Chinese culture plays in online communication and how it influence social capital.

Workshops

Quality Control Procedures for the Scoring and Rating of Tests in Different Environments and Administration Modes

Avi Allalouf

National Institute for Testing and Evaluation, Israel

e-mail: avi@nite.org.il

Quality control (QC) procedures are required in order to monitor the scoring process. Failure to establish and implement such procedures can lead to inaccurate scores being calculated with potentially serious consequences, such as a qualified candidate not being accepted to a university or place of employment. The workshop will provide examples from real-life contexts, and participants will be given hands-on practice in applying QC procedures in scoring and rating tests in different environments and administration modes. The main topics to be presented are: (1) QC procedures for large-scale assessments with large and stable cohorts – usually in paper & pencil mode; (2) QC procedures for scores on tests administered to small population groups on multiple administration dates, (Continuous Administration Mode) – usually computer- and Internet-based; (3) monitoring the quality of performance assessment raters who conduct offline and online scoring; and (4) procedures to prevent and detect cheating.

Comparative Judgments as an Alternative to Rating Scales: Designing and Scoring Forced-Choice Questionnaires

Anna Brown

University of Kent, UK

e-mail: a.a.brown@kent.ac.uk

To counter response distortions associated with the use of rating scales in personality and similar assessments, test items may be presented in so-called 'forced-choice' formats. Respondents may be asked to rank-order a number of items, or distribute a fixed number of points between several items – therefore they are forced to make a choice. Until recently, basic classical scoring methods were applied to such formats, leading to scores relative to the person's mean (ipsative scores). While interpretable in intra-individual assessments, ipsative scores are problematic when used for inter-individual comparisons. Recent advances in multidimensional item response theory and estimation methods enabled rapid development of item response models for comparative data, including the Multi-Unidimensional Pairwise Preference Model (Stark, Chernyshenko & Drasgow, 2005), and Thurstonian IRT model (Brown & Maydeu-Olivares, 2011).

This workshop will introduce participants to state-of-the-art methods for designing and scoring forced-choice questionnaires, combining theory with hands-on exercises. We will provide practical guidance on how to code, analyse and score forced-choice responses under Thurstonian IRT modelling, using Mplus (Muthén & Muthén, 1998-2012) and R (R Development Core Team, 2008). The learning objective of the workshop is to equip researchers with skills needed to improve the existing or develop new forced-choice questionnaires.Participants should be familiar with item response theory and factor analysis.

Multigroup Modeling with Cross-National Data: Applications, Issues, and Complexities

Barbara B. Byrne

University of Ottawa, Canada

e-mail: bmbyrne@comcast.ne

This workshop addresses the topicsof measurement equivalence and hierarchical data structure in multigroup comparisons across samples representing different countries. Participants are "walked through" applications based on the EQS and Mplusprograms that include testing: (a) measurement and structural equivalence of assessment scales, (b) latent mean differences, and (c) multilevel modeling. Issues and complexities that can impact cross-national comparisons are described and illustrated. To gain the most from this workshop, participants should have some understanding of basic SEM concepts.

Observed-Score Test Equating: An Overview

Alina von Davier

Educational Testing Service, USA

e-mail: avondavier@ets.org

Test equating methods are used to produce scores that are comparable across different test forms. The observed-score equating (OSE) framework is a unified approach to test equating based on a flexible family of equipercentile-like equating functions that contains the linear equating function as a special case. The framework is based on the kernel method (KE) of test equating. Any OSE is viewed as having five steps: 1) pre-smoothing; 2) estimation of the score probabilities on the target population; 3) continuization; 4) computing the equating function; 5) computing the standard error of equating and related accuracy measures. The OSE framework brings together these steps into an organized whole rather than treating them as disparate problems. The framework exploits pre-smoothing by fitting statistical models (log-linear models) to score data, and incorporates it into step 5) above. The OSE includes new tools for comparing two or more equating functions and to rationally choose between them.

In this session, theoretical issues will be considered along with numerical examples and software demonstration using real data. The session will provide an overview of the observed-score equating methods, their assumptions, and the relevant data collection designs. The book "The Kernel Method of Test Equating" of von Davier, Holland, and Thayer (2004) is the basis of this training session. The LOGLIN/KE Software is described. A demo of equating with an equivalent-groups design with real data will be provided.

Assessing 21st Centurys Skills

Kurt F. Geisinger
University of Nebraska, USA

e-mail: kgeisinger@buros.org

The changing nature of skills needed for school and job success are sometimes referred to as 21st century skills. These skills de-emphasize memory, for example, because many items that had traditionally been memorized are now easily electronically retrievable. Similarly, with the easy access to information, one must know the value of information; some information provided on the web is simply not accurate, for example.

Some have collapsed the proposed 21st Century skills into three broad clusters as shown below:

Cognitive skills: nonroutine problem solving, critical thinking, systems thinking;

Interpersonal skills: complex communication, social skills including collaboration, teamwork, cultural sensitivity, dealing with diversity;

Intrapersonal skills: self-management, time management, self-development, self-regulation, adaptability, executive functioning; and

Technical skills: Researching and information fluency skills.

This workshop will discuss ways that such skills can and have been assessed and look at future attempts to improve such measures. The use of performance assessments as well as more traditional assessments will be discussed.

Item Response Theory: Concepts, Models, and Applications

Ronald K. Hambleton
University of Massachusetts at Amherst, USA

e-mail: rkh@educ.umass.edu

Many testing agencies and researchers would like to use item response theory (IRT) models for developing, scoring, and equating aptitude, achievement, and personality tests, studying bias, and reporting test scores. 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 half-day workshop, we will survey the following topics: (1) Shortcomings of classical test theory that have inspired the development of IRT models, (2) specific IRT models for fitting binary and polytomously-scored data and model assumptions, (3) basics of item and ability parameter estimation, (4) graphical and statistical approaches for assessing model fit, (5) to IRT software, (6) test development using item and test information, (7) equating of test scores, (8) studying item level bias, computer-adaptive testing designs, and (9) score reporting. Because of the limited time available to cover these topics, and most topics will be touched on only briefly, we will provide a bibliography to facilitate follow-up reading. Also, and again because of limited time, and the difficulties of having attendees run any statistical analyses during the short workshop time, demonstrations of running IRT software will be included in the presentation, and short tests will be administered to help participants monitor their progress with the workshop content.

Test Adaptation: The Strife for Equivalence

Dragos Iliescu

SNSPA University, Bucharest, Roman

e-mail: dragos.iliescu@testcentral.ro

The workshop will focus on a number of topics related to test adaptation. The will focus on terminology, legal constraints, and the general test adaptation sequence. The bulk of the workshop will be dedicated to the problem of equivalence, focusing on linguistic equivalence, cultural equivalence and psychological equivalence (construct and measurement equivalence). Each of these points will be illustrated with case studies. Finally, caveats in the test adaptation process are underlined, by discussing sources of bias affecting test adaptation (Hambleton et al., 2004). The sources of bias will be discussed under three headings: (a) Cultural & language differences, (b) Technical aspects (design of the test, design of the adaptation process), and (c) Interpretation of test results. The workshop is not oriented towards statistical procedures, but rather towards cultural localization issues.

Latent Class Analysis: Applications to Test Data

Bruno D. Zumbo

University of British Columbia, Canada

e-mail: bruno.zumbo@ubc.ca

Nearly all contemporary psychometric models make use of latent variables, of one form or another. An important consideration in using these models is the performance of the items in potentially heterogeneous populations of respondents. Many claims about the validity of our measures, and the inferences we make there from, are based on the premise that individuals interpret and respond to sets of items in a consistent manner such that the measurement model parameters are equivalently applicable to all people irrespective of any differences among them in our target population. The purpose of this workshop is to introduce latent class analysis in the context of the analysis of test data. The use of latent variable mixture models will be introduced and demonstrated in the context of examining the extent to which a sample is homogeneous with respect to a specified unidimensional model for categorical data and identify potential sources of sample heterogeneity. In addition to test level analyses, we will also introduce a new set of latent class methods recently introduced by the author to investigate heterogeneity at the item level – a new latent class item bias technique. The workshop is structured to focus on the fundamentals of the methods and demonstrate the techniques with real test data. Prior basic knowledge of structural equation modeling and confirmatory factor analysis, in particular, will be assumed.

Index of Contributors

Last minute addition to the program

Detecting Misfitting Response Patters in Educational Testing: An Empirical Application

Rob R. Meijer and Jorge N.Tendeiro
University of Groningen, Netherlands
e-mail: r.r.meijer@rug.nl

Person-fit indices were applied to detect inconsistent patterns of correct/incorrect answers to test questions (items) on a high-stakes educational test. We were particularly interested in (a) the presence of aberrant item score patterns and their configurations of item scores (b) the relation of inconsistent answer patterns to given subgroups of respondents. We show that persons that were consistently flagged across subtests had relatively low scores, which may be the result of extensive guessing. We found no relation of inconsistent behavior between males/females and for most subgroups. However, we found significant differences in mean person-fit scores for one subgroup when compared to other subgroups. This particular subgroup has a large proportion of non-native English speakers. We argue that person-fit indices should be used to routinely monitor test behavior.

Notes:
* Withdrawn presentations
** Single Papers

Sponsors

Join the Assessio network!

Assessio is the major test publishing company in the Nordic region. The company was started in 1954 by the Swedish Psychological Society and has a long history of developing research-based tests for selection and development of personnel. We are privately held since 1991.

Assessio's focus today is international expansion and global distribution. Our tests are available through our partner network in over 30 countries around the world. We offer a complete web-based portfolio to local test vendors and consultancy firms:

- Screening – **ServiceFirst** and **MINT**
- Big Five Personality – **MAP**
- General Mental Ability – **Matrigma**

To learn more about Assessio, our tests and business opportunities, visit our booth in the exhibition area.

info@assessio.se
www.assessio.com

ASSESSIO

smart. valid. preferred.

Listening. Learning. Leading.®

UPV / EHU

UDA IKASTAROAK
CURSOS DE VERANO

Consejo General de Colegios Oficiales de Psicólogos

BUROS

CENTER FOR TESTING

WWW.BUROS.ORG

Improving the science and practice of testing and assessment

PSYCHOMETRIC CONSULTING

- Expert psychometric services
- Independent verification processes
- Improved quality of proprietary testing programs

ASSESSMENT LITERACY

- Instructional and educational resources
- Informed selection, development, and use of tests
- Promotes responsible test use

TEST REVIEWS & INFORMATION

- Authoritative reference materials
- Mental Measurements Yearbooks
- Tests in Print
- Pruebas Publicadas en Español

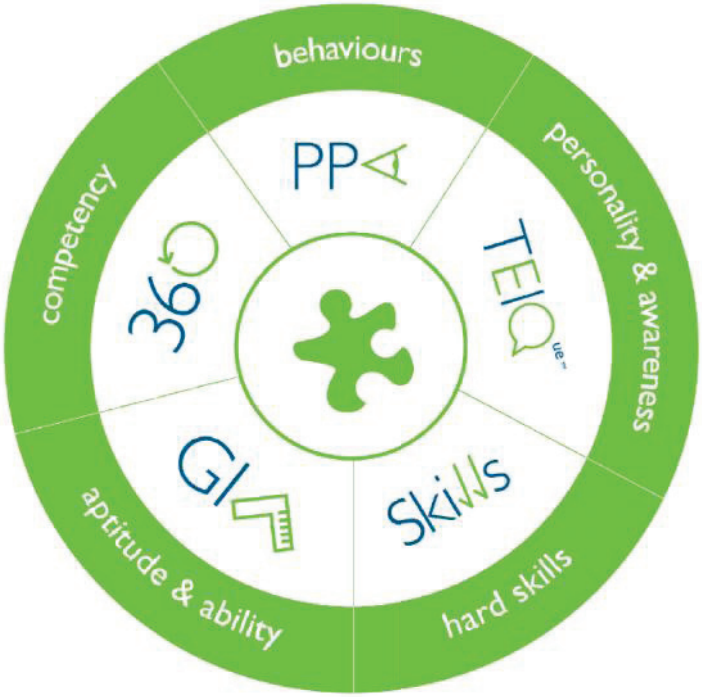
University of Nebraska Lincoln | COLLEGE OF EDUCATION & HUMAN SCIENCES

thomas
your business. your people. but better

Transform the performance of your teams and individuals

Thomas International provides people assessments which empower business leaders to transform the performance of their teams and individuals – and deliver an immediate impact on their organisation.

Our assessments work together to give businesses full insight into what makes people successful – their behaviours, personality, aptitude and ability, competency and hard skills.



TEIQue

Emotional intelligence gives your people the edge they need to perform in today's complex business environment. The TEIQue tells you how well your people:

- Understand their emotions
- React to pressure
- Manage relationships

Skills

Skills is a range of tests and training modules designed to:

- Measure skill levels
- Ensure you're optimising the skills your staff have
- Help your staff develop new skills

GIA

GIA measures a person's mental ability, giving you a prediction of their potential to grasp a new role or respond to training by answering questions such as:

- Can they think on their feet?
- Could they be a high flyer?
- Are they problem solvers?

360

360 provides a structured framework for:

- Identifying performance gaps
- Developing self-awareness, confidence and motivation
- Understanding how to improve personal effectiveness

PP

In just 8 minutes PPA provides an accurate insight into how people behave at work, answering questions such as:

- What are their strengths and limitations?
- How do they communicate?
- Are they self-starters?
- What motivates them?

For more information, visit or call us:

www.thomasinternational.net
+44 (0) 1628 475 366

GRADUATE MANAGEMENT
ADMISSION COUNCIL

CEB | SHL Talent Measurement

Applying Science to the Measurement of Talent

Data gathered from testing can be linked to key metrics for an organisation to help demonstrate the overall impact of talent measurement. It is critical, therefore, to ensure that objective assessments can be applied throughout the whole employee lifecycle.

SHL Talent Measurement, part of the CEB talent management portfolio, help to predict performance through scientifically respected assessments. Among our 'firsts' include:

- Online IRT-based cognitive employment tests
- Video-based multi-media situational judgement tests
- Online verification of unsupervised cognitive testing
- Multidimensional IRT scoring for forced-choice personality format items
- Computer-adaptive personality assessments
- Talent Analytics

Along with continuous innovation in our assessment technologies, our scientists are regular contributors to national and international scientific conferences, leading peer-reviewed journals and other publications.

For the latest on the trends on the use of objective assessments in the work-place, download our 2014 Global Assessment Trends Report at **ceb.shl.com/gatr**

Get in touch at www.ceb.shl.com | info@shlcom

The 9th International Test Commission Conference (ITC) took place at the Miramar Palace in San Sebastian, Spain, between the 2nd and 5th of July, 2014. The Conference was titled, "Global and Local Challenges for Best Practices in Assessment."

The International Test Commission, ITC (www.intestcom.org), is an association of national psychological associations, test commissions, publishers, and other organizations, as well as individuals who are committed to the promotion of effective testing and assessment policies and to the proper development, evaluation, and uses of educational and psychological instruments. The ITC facilitates the exchange of information among members and stimulates their cooperation on problems related to the construction, distribution, and uses of psychological and educational tests and other psychodiagnostic tools.

This volume contains the Abstracts of the contributions presented at the 9th International Test Commission Conference. The four themes of the Conference were closely linked to the goals of the ITC:

- Challenges and Opportunities in International Assessment
- Application of New Technologies and New Psychometric Models in Testing
- Standards and Guidelines for Best Testing Practices
- Testing in Multilingual and Multicultural Contexts

9 788498 609974