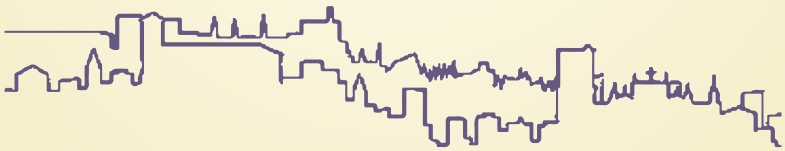
CONFERENCE PROGRAM

ITC CONFERENCE

02 · 05 JULY 2024



G R A N A D A

INTERNATIONAL TEST COMMISSION

CONTENTS

duolingo english test

englishtest.duolingo.com/edu

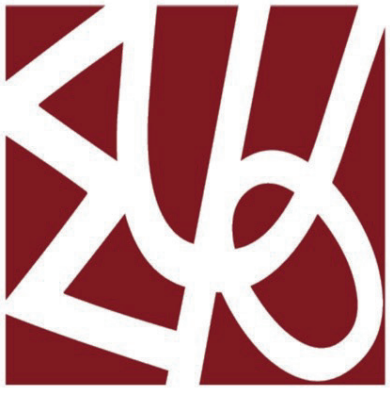
The future of testing. Here today.

Expand your applicant pool
Tap into a diverse pool of candidates from 210+ countries and territories of origin, who have taken the Duolingo English Test because of its radical accessibility.

Built on the latest assessment science
The Duolingo English Test is a computer adaptive test powered by rigorous research and AI. Results are highly correlated with other assessments, such as the TOEFL and the IELTS.

Innovative test security
Industry-leading security protocols, individual test proctoring, and computer adaptive technology help prevent fraud and cheating and ensure results you can trust.

Convenient results dashboard
Access candidates' certificates, video interviews, and writing samples through a free and secure dashboard. Quickly and easily view applicant data with filtering tools.



UMassAmherst

College of Education

Center for Educational Assessment

We add life to a lifetime of learning.

Come and see us at Stall 3.

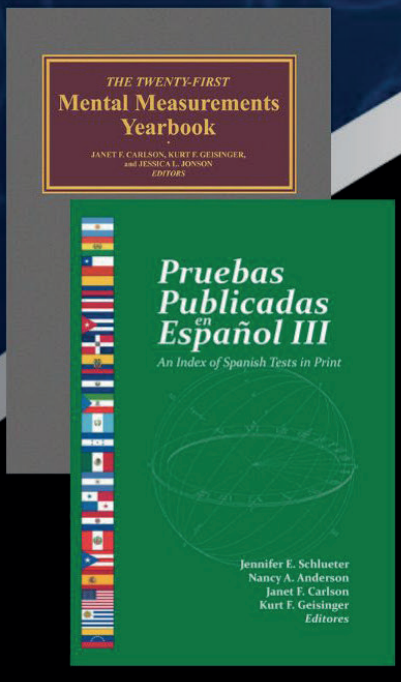
Pearson

BUROS

CENTER FOR TESTING

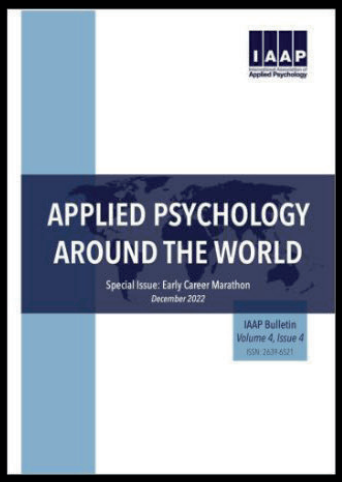
TEST REVIEWS • ASSESSMENT LITERACY
PSYCHOMETRIC CONSULTING

WWW.BUROS.ORG



ENGLISH AND SPANISH
TEST SELECTION RESOURCES

PROFESSIONAL DEVELOPMENT
RESOURCES
https://buros.org/assessment



APPLIED
RESEARCH

More than a publisher.

1949–2024

The cornerstones of our work:

- Locations in 16 countries
- 6.500 authors
- 500 training and qualification courses
- 2.300 tests available in 24 languages
- 2,400 books available in 12 languages
- 140 new book publications per year
- 39 journals
- 145 exhibitions per year at conferences and scientific meetings
- 500 professionals worldwide

Proud to be a Platinum sponsor of ITC

www.hogrefe.com

hogrefe

*ets research institute

Reimagining measurement to drive impact

As leaders in assessment and measurement, we are reshaping paradigms to drive human progress.

Learn more at ets.org/research

cApStAn
linguistic quality control

LANGUAGE SOLUTIONS FOR LINGUISTIC AND CULTURAL COMPARABILITY OF YOUR HIGH-STAKES TESTS

Translatability Assessment | DEI Consultancy | Translation Workflows
Machine Translation Post-Editing | Expert Linguistic Verification

bizdev@capstan.be
www.capstan.be

the british
psychological society
psychological testing centre

The British Psychological Society's Psychological Testing Centre

The UK's leading national organisation for setting standards in psychological testing.

We provide information and services relating to standards in tests and testing for test takers, test users, test developers and the general public.

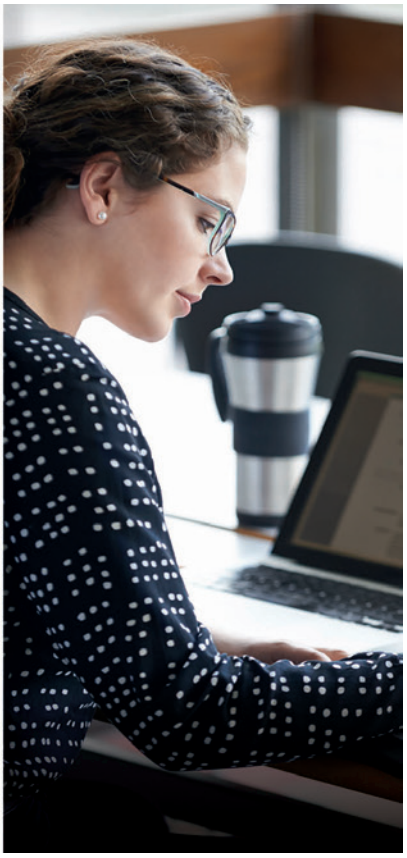
We have developed qualification standards for professionals using tests in the Occupational, Forensic and Educational domains which define safe and responsible practice.

Our Register of Qualified Test Users (RQTU) enables qualified professionals to be found.

Expert testing knowledge within the BPS has enabled the PTC to publish over 160 test reviews – so test users can identify suitable tests.

Our publication *Assessment & Development Matters* (ADM) and other information can be accessed via the website **www.psychtesting.org.uk**

PSYCHOLOGICAL TESTING CENTRE



SPONSORS

GALA DINNER SPONSOR // ALHAMBRA VISIT SPONSOR

duolingo english test

UMassAmherst
College of Education
Center for Educational Assessment

DIAMOND SPONSORS

Pearson

BUROS CENTER FOR TESTING

hogrefe

GOLD SPONSORS

ets

cApStAn linguistic quality control

the british psychological society
promoting excellence in psychology

SILVER SPONSORS

AAMC

NBME

ACS VENTURES
Bridging Theory & Practice

ASSESSFIRST

edCount LLC
because all students count

Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego

American Psychological Association

caveon

HumRRO
Human Resources Research Organization

IEA
Researching education, improving learning

eMetric

Graduate Management Admission Council

ncme
National Council on Measurement in Education

ACER

Commsmultilingual
Dependable • Flexible • Defensible

COMMITTEES

LOCAL ORGANIZING COMMITTEE (LOC):

PRESIDENCY

José-Luis Padilla
Full Professor at the Dept. of Methodology of Behavioral Sciences, University of Granada (Spain), and member of the Mind, Brain, and Behavior Research Centre (CIMCYC).

Luis-Manuel Lozano
Associate Professor at the Dept. of Methodology of Behavioral Sciences at the University of Granada (Spain), and member of the Mind, Brain, and Behavior Research Centre (CIMCYC).

Isabel Benítez
Associate Professor at the Dept. of Methodology of Behavioral Sciences at the University of Granada (Spain), and member of the Mind, Brain and Behavior Research Centre (CIMCYC).

MEMBERS

Andrés González
Associate Professor at the Dept. of Methodology of Behavioral Sciences at the University of Granada (Spain), and member of the Mind, Brain and Behavior Research Centre (CIMCYC).

Ignacio Martín
Full Professor and Director of the Dept. of Methodology of Behavioral Sciences at the University of Granada (Spain).

María-Carmen Aguilar-Luzón
Associate Professor and Director of the Dept. of Social Psychology at the University of Granada (Spain) and member of the Mind, Brain and Behavior Research Centre (CIMCYC). President of the Spanish Association of Environmental Psychology (PSICAMB).

Mariela Bustos Ortega
Postdoctoral Researcher at in the Dept. of Methodology for Behavioral Sciences at the University of Granada (Spain), and member of the Mind, Brain and Behavior Research Centre (CIMCYC).

María-Carmen Navarro-González
PhD Student in the Dept. of Methodology of Behavioral Sciences, and member of the Mind, Brain and Behavior Research Centre (CIMCYC).

Rocío Vizcaíno-Cuenca
PhD Student in the Dept. of Methodology for Behavioral Sciences at the University of Granada (Spain), and member of the Mind, Brain and Behavior Research Centre (CIMCYC).

Belén Carrascal-Caputto
PhD Student in the Dept. of Methodology for Behavioral Sciences at the University of Granada (Spain), and member of the Mind, Brain and Behavior Research Centre (CIMCYC).

Juan F. Luesia-Lahoz
Research Assistant in the area of the Methodology of Behavioral Sciences at the Loyola University Andalusia (Spain).

Celia Serrano-Montilla
Assistant Professor at the National University of Distance Education (UNED) (Spain)

Álvaro Postigo
Assistant Professor at the University of Oviedo (Spain).

Jorge Torres Marín
Assistant Professor at the Dept. of Developmental and Educational Psychology at the University of Granada (Spain).

David Sánchez Casasola
PhD Student at the Dept. of Methodology for Behavioral Sciences at the University of Granada (Spain).

Albert Sesé
Full Professor at the Department of Psychology of the Balearic Islands University (Spain), former president of the Stress and Anxiety Research Society (STAR),and the current President of the European Association of Methodology (EAM).

Ana Hernández
Associate professor at the University of Valencia (Spain), member of the executive committees of AEMCCO (Spanish Association of Methodology of Behavioral Sciences), EAM (European Association of Methodology), and of the Board of Assessment of the EFPA (European Federation of Psychological Associations). She is vice-president of the AEMCCO (Spanish Association of Metyhodology of Behavioral Sciences)

Nekane Balluerka
Full Professor of Methodology of Behavioral Sciences at the Faculty of Psychology of the University of the Basque Country (Spain). She is president of the Spanish Association of Methodology of Behavioral Sciences (AEMCCO), and member of the Executive Committee of the European Association of Methodology (EAM). She was Rector of the University of the Basque Country (2017-2021).

SCIENTIFIC COMMITEE:

Miguel Sorrel
Universidad Autónoma de Madrid (España).

Dorothée Behr
GESIS Leibniz Institute for the Social Sciences (Alemania).

Kadriye Ercikan
ETS- Educational Testing Service (EE.UU).

Amina Abubakar
Pwani University (Kenia).

Steve Dept
cApStAn (EE.UU y Bélgica).

María Dolores Hidalgo
Universidad de Murcia (España).

INTERNATIONAL ADVISORY:

José Muñiz
Universidad Nebrija (España).

Felipe Valentini
Universidade São Francisco (Brasil).

Alejandra Domínguez
Universidad Iberoamericana Ciudad de México (México).

Bruno Zumbo
University of British Columbia (Canadá).

Priyanka Sharma
Australian Council for Educational Research (India).

Ahmad Fasfous
Bethlehem University (Palestina).

Ruben Ledesma
IPSIBAT, Instituto de Psicología Básica, Aplicada y Tecnología, CONICET y Universidad Nacional de Mar del Plata (Argentina).

Madeline Xi Xiamoing
Hong Kong Examinations and Assessment Authority (Hong Kong).

MAIN THEMES AND THEMATIC LINES:

Thematic line 1: Challenges and solutions for translation.

Thematic line 2: Adapting evaluation instruments to minority languages and cultures.

Thematic line 3: Survey research, psychometrics, and psychological assessment: Shared challenges and solutions.

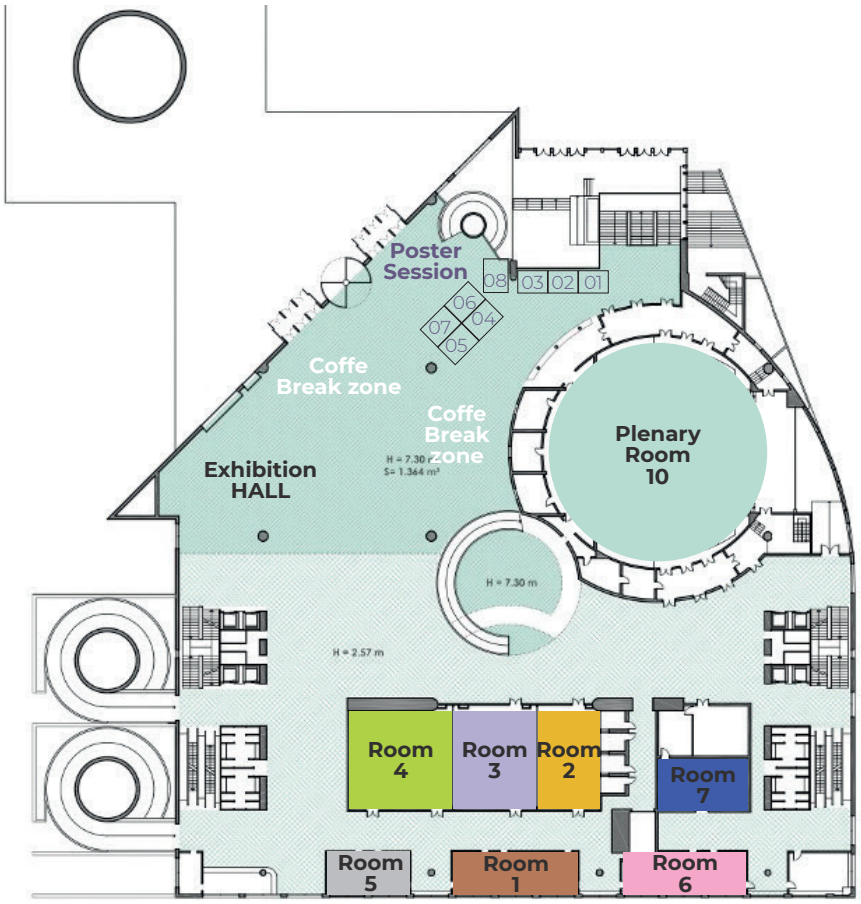
TOPICS:

Translation/adaptation of tests, psychological assessment instruments, and survey questionnaires.
Construct or concept equivalence.
Testing equivalence by psychometrics methods.
Identifying biases by qualitative or quantitative methods.
Validity theory in testing, psychological assessment and survey research.
Quantitative, qualitative, and mixed validation methods.
Validity and fairness in cross-cultural testing, psychological assessment and survey research.
Computational developments for social science research in cross-cultural testing.
Artificial Intelligence in testing, psychological assessment and survey research.
Innovations in test development.
International assessment.
Psychometric modeling.

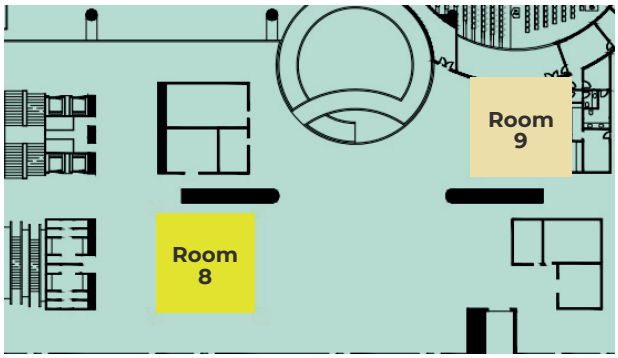
CONFERENCE VENUE

Conference venue at Palacio de Congresos, Granada, Spain.
Paseo del Violón, s/n. 18006 Granada

FIRST FLOOR



SECOND FLOOR



PROGRAMME OVERVIEW

	TUESDAY
9:00am-12:00pm	**Short course** (Morning session)
12:30pm-3:30pm	**Short course** (Afternoon session)
4:00pm-7:00pm	**Short course** (Evening session)
7:30pm	**Welcome reception**

	WEDNESDAY	THURSDAY	FRIDAY
8:30am-9:45am	Session 1	Session 5	Session 9
9:55am-10:55am	Conference Opening and ITC Presidential Address.	Invited Panel	Keynote
10:55am-11:45am	Coffee break	Coffee break	Coffee break
11:55am-11:45am	Poster session 1	Poster session 3	Poster session 5
11:45am-1:00pm	Keynote	Session 6	Session 10
1:00pm-2:00pm	Conference Lunch & Graduate Student Committee Meeting	Conference Lunch & Early Career Scholars Meeting	Thomas Oakland Award and Closing Ceremony
2:00pm-3:15pm	Session 2	Session 7	
3:25pm-4:40pm	Session 3	Session 8	ITC Council Meeting
4:40pm-5:30pm	Coffee break	Coffee break	
4:40pm-5:30pm	Poster session 2	Poster session 4	
5:30pm-6:30pm	Session 4	ITC General Meeting	
7:15pm-8:00pm	Alhambra Tour		
8:00pm-8:30pm			
8:30pm		Gala Dinner	

TUESDAY JULY 2nd

9:00am - 12:00pm **SHORT COURSES** Morning session

- **ROOM 9**
  An Introduction to the Theory of Standard Setting
  Mark D. Reckase *(College of Education at Michigan State University)*
- **ROOM 3**
  Response Bias in Self-Report Measurements
  Ricardo Primi *(Graduate School of Psychology Assessment at the University of São Francisco in Campinas, Brazil)* **& Felipe Valentini** *(Graduate School of Psychology Assessment at the University of São Francisco in Campinas, Brazil)*
- **ROOM 2**
  Best Practices for Artificial-Intelligence Scoring of Constructed Responses
  Daniel F. McCaffrey (ETS)

12:30pm - 3:00pm **SHORT COURSES** Afternoon session

- **ROOM 3**
  Questionnaire translation/adaptation: On ensuring equivalence in cross-cultural research
  Dorothée Behr *(GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany)* **& Brita Dorer** *(GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany)*
- **ROOM 2**
  Cognitive Diagnosis Modeling: Introduction, Recent Developments, and Implementation in R.
  Jimmy de la Torre *(Faculty of Education at the University of Hong Kong)* **& Pablo Nájera** *(Department of Psychology, Universidad Pontificia Comillas)*
- **ROOM 9**
  Sequence Mining Methods on Process Data in Large-Scale Assessments
  Qiwei He *(Georgetown University)*, **Esther Ulitzsch** *(Leibniz Institute for Science and Mathematics Education in Kiel, Germany)* **& Bernard Veldkamp** *(University of Twente, the Netherlands)*

4:00pm - 7:00pm **SHORT COURSES** Evening session

- **ROOM 3**
  Introduction to AI-based Automated Item Generation and Scoring Practices
  Alina Von Davier *(Chief of Assessment, Duolingo)* **& Duanli Yan** *(Educational Testing Service)*
- **ROOM 2**
  Solving the problems of ipsative data: Designing and scoring forced-choice and other questionnaires in comparative format
  Anna Brown *(University of Kent)*
- **ROOM 9**
  A Critical Quantitative measurement perspective: MIMIC models to identify and remediate racial (and other) forms of bias
  Matthew Diemer *(University of Michigan)*

7:30pm **WELCOME RECEPTION** (Terrace of the Granada Conference Palace)

Stephen G. Sireci, *University of Massachusetts Amherst, USA, ITC President*

José Luis Padilla García, *University of Granada, Spain, Co-President of the Local Organizing Committee ITC Conference 2024*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

WEDNESDAY JULY 3rd

9.00am-9.45am Session 1.1. **TOPIC** Innovations in test development
Room 1
Chair: Emerik Kubiak *(AssessFirst)*

Evaluating and Evolving Cognitive Assessments in the Age of Large Language Models
Emeric Kubiak *(AssessFirst)*

A Cross-Cultural Test explained by a Scientific Dialectical Discourse
Claudia Gusso *(International Test Commission (ITC))*

Measurement invariance/equivalence of the Index of Psychological Well-being at Work Across Black and White South African employees
Gina Görgens-Ekermans *(Stellenbosch University)*

Measuring Personality in a Changing World: Psychometric Analyses of the BFI-2 across Offline and Online Situations
Dora Leander Tinhof *(Bielefeld University)*

Children's Resilience Markers: Initial Studies for Age Range Expansion
Karina da Silva Oliveira *(Universidade São Francisco)*

8.30am-9.45am Session 1.2. **TOPIC** International Assessment
Room 2
Chair: Yan Zhang *(University of Auckland)*

An Examination of School-Based and Work-Based Assessment Practices at China's Tertiary TVET Institutes
Yan Zhang *(University of Auckland)*

Item Quality for Cognitive Assessments in Low-to-Middle Income Countries: Evidence from the Ethiopia Young Lives Data
Winifred Wilberforce *(Ohio State University)*

Towards Multi-Cultural Appropriate Assessments in South Africa: Challenges with the Test Classification Process
Justin August *(Health Professions Council of SA)*

Pēhea Tōku Haerenga? A Māori Self-Assessment Measure for Aging and End of Life
Melissa Carey *(University of Southern Queensland)*

8.30am-9.45am Session 1.3. **SYMPOSIUM** Examinee Engagement and Affect in Low-Stakes Testing Contexts: Influences and Personalized Interventions
Topic: Validity theory in testing, psychological assessment and survey research
Room 3
Chair: Sara Finney *(James Madison University)*

The Question-Behavior Effect in Low-Stakes Testing Contexts: How Many Questions are Needed to Prompt Good Test-Taking Effort?
Sara Finney *(James Madison University)*

Perceived Normativity Of Giving Effort On Low-Stakes Tests: Measures And Relations With Examinee Effort And Test Performance
Dena Pastor *(James Madison University)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

A Personalized Intervention Study On Decreasing Not-Fully-Effortful Responses During Low-Stakes Mathematics Assessment
Burcu Arslan *(Educational Testing Service Global)*

Exploring the Affective Impact of Immediate Feedback in Low-Stakes Testing: The Role of Student Performance and Visual Feedback Design
Livia Kuklick *(IPN Kiel, Germany)*

8.30am-9.45am Session 1.4. **TOPIC** Translation of tests, psychological assessment instruments and survey questionnaire
Room 4
Chair: Paul Shrell-Fox *(Efrata Academic College, PsychTech)*

Language Cultural Considerations and Outcomes in the Translation of WRAT5 to Hebrew
Paul Shrell-Fox *(Efrata Academic College, PsychTech)*

Watch out for that item: Considerations on the linguistic adaptation of the Comprehensive Aphasia Test into Malay
Giuditta Smith *(University of East Anglia)*

Psychological Assessment in Brazil: Perspectives and Challenges
Solange Muglia Wechsler *(Pontificial Catholic University of Campinas)*

Challenges of Psychological Assessment: The case of Chile
Marcela Rodríguez-Cancino *(Universidad de La Frontera)*

8.30am-9.45am Session 1.5. **TOPIC** Psychometric modeling
Room 5
Chair: E. Cihat Corbaci *(Sinop University)*

Latent growth analysis of serial eye-fixation indicators for multiple-choice test items
E. Cihat Corbaci *(Sinop University)*

Psychometric evaluation of the environmental knowledge scale in TIMSS 2019
Purya Baghaei *(IEA-Hamburg)*

A Latent Profile Analysis of Examinees' Multiple-Choice Item Processing Behavior Using Segment-Specific Eye-Fixation Metrics
Derya Akbas *(Aydın Adnan Menderes University)*

Willing and Able to Fake: A Flexible Item Response Modeling Framework for Applicant Faking Measurement (Psychometric modeling)
Siwei Peng *(Jiangxi Normal University)*

Dynamic Structural Equation Modeling of Daily Happiness and Stress Data
Esra Sözer Boz *(Bartin University/Turkey)*

8.30am-9.45am Session 1.6. **SYMPOSIUM** Changes in Collegiate Admissions Policies and Procedures. Topic: Validity theory in testing, psychological assessment and survey research
Room 6
Chair: Kurt Geisinger *(Buros Center for Testing, University of Nebraska-Lincoln)*

Changes in Admission Tessting Practices in Colleges and Universities
Kurt Geisinger *(United States)*

Global Perspectives on Higher Education Admissions: Navigating Access, Diversity, and Equity Challenges (International assessment)
Maria Elena Oliveri *(Buros Center for Testing)*

Test optional policies: trends and impact - to date (Validity and fairness in cross-cultural testing, psychological assessment and survey research)
Wayne Camara *(LSAC)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

8.30am-9.45am

Session 1.8. **TOPIC** Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research
Room 8
Chair: Hongwen Guo *(ETS)*

Integrating Process Data With Response Data for Cross-Cultural Comparability and Data Insights
Hongwen Guo *(ETS)*

Threats To Validity and Fairness of Testing: How Different Demographic and Cross-Cultural Groups Experience Testing Differently
Stephen Cuppello *(Thomas International)*

Institutional Diversity and its Implications for Assessments and Outcomes
Indrani Bhaduri *(NCERT)*

Mixed Item Type Strategies to Support Culturally Relevant Cross-Cultural Measurement
Fernanda Gandara *(Room to Read)*

9.00am-9.45am

Session 1.9. **TOPIC** Validity and fairness in cross-cultural testing, psychological assessment and survey research
Room 9
Chair: Qilong Zhang *(United Arab Emirates University)*

Self-assessment scale of emotional labour for early childhood teachers: A context-centered approach
Qilong Zhang *(United Arab Emirates University/United Arab Emirates)*

Cognitive and Emotional Processing of Culturally Responsive Test Items
Chris Patterson *(University of Iowa)*

Fortifying the Human Firewall: Development and Validation of a Personality-Based Organizational Cybersecurity Risk Assessment Framework
Aishwarya Jaiswal *(Mercer | Mettl)*

9:55am-10:55am

Conference Opening and ITC Presidential Address
Plenary Room 10

Welcome by Local Organizing Committee (LOC). *José Luis Padilla, University of Granada. Co-President of the Local Organizing Committee ITC Conference 2024*

Welcome by a representative from the local Government

Welcome by the ITC President. *Stephen G. Sireci, University of Massachusetts Amherst.*

ITC Presidential Address. The Failure and Future of Psychometrics
Stephen G. Sireci, *University of Massachusetts Amherst, USA, ITC President*

10:55am-11:45am

Coffee break

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

10:55am-11:45am **Poster session 1**

Room: Exhibition Hall

TOPIC Artificial Intelligence in testing, psychological assessment and survey research

1 Enhancing ADHD Screening in Children: Integrating Machine Learning with Item Response Theory.
Pedro Loures Alzamora, Derick Oliveira, Camila Nicola, Victoria Oliveira, Laura Ludgero, Ana Paula Couto Silva, Gisele Pappa, Marco Romano-Silva, Alexandre Serpa, Wagner Meira Jr, Débora Marques Miranda

2 AI for Psychometrics: Validating Machine Learning Models in Measuring Emotional Intelligence with Eye-Tracking Techniques.
Wei Wang, Liat Koffler, Chapman Lindgren, Max Lobel, Amanda Murphy, Qiwen Tong, Kemar Pickering

3 Exploring the Reliability of Audio Signals in Video Interviews for the Automatic Prediction of Psychological Characteristics.
Borja Artiñano, David Aguado, Pablo Garcia, Sara Estirado

4 Distilling vector space models for psychoeducational assessment: honing semantic indicators in automated summary evaluation.
José Ángel Martínez-Huertas, Guillermo Jorge Botana, Ricardo Olmos, José A León

5 An Examination of Racial Bias in Artificial Intelligence When Conducting Neurological and Learning Assessments.
Joseph Kush, Inna Vaisleib

TOPIC International assessment

6 Measurement invariance of a general cognitive performance measure across 26 European countries and Israel.
Adrián García Mollá, José Manuel Tomás, Amparo Oliver, Zaira Torres, Irene Fernández

7 Quantitative Literacy proficiency in Mathematics students: a cross-domain diagnostic exploration.
Tatiana Sango, Sanet Steyn

TOPIC Psychometric modeling

8 RaterLynx: A Shiny App for Incomplete Rating Designs of Rater-mediated Assessments.
Angel Arias

9 Comparative Analysis of Psychometric Models for Testlet-Structured Assessments.
Carlos David Diaz Lopez, Joaquín Caso Niebla, Coral González Barbera

TOPIC Innovations in test development

10 Research & Invention Methodology to generate a Cross-Cultural Test.
Claudia Gusso.

11 Stability Analysis of Estimation of Piecewise Linear ICCs.
Gen Hori, Sayaka Arai

12 Construction of a Situational Judgment Test for the selection of prison officer candidates: from job analysis to item production.
Gucek Richard, Loarer Even, Terriot Katia

13 Relationship between interests differentiation and social-emotional skills in Brazilian students: A new interest differentiation index.
Gustavo Henrique Martins, Filip De Fruyt, Ana Carla Crispim, Joyce Scheirlinckx

14 Advanced Care Planning in Mental Health: From Systematic Review to Instrument Development.
Chao Zhang, Maite Barrios, Ángela Berrío, Juana Gómez-Benito, Georgina Guilera

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

15 Student Evaluation in a Diverse Educational Environment: India's Holistic Progress Card.
Paul Borysewicz, Aakanksha Bhatia, Neeraj Venkatamaran, Jonas Bertling, Indrani Bhaduri

16 Damage to Honor Scale: A Forensic Psychological Contribution to the Evaluation of Moral Damage.
Yessica Daniela González Berriel, Alejandra del Carmen Domínguez-Espinosa, Marina Flores-Camargo

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

11:45am-1:00pm **KEYNOTE Embracing the Future of Remote Proctoring: Ensuring Test Integrity and Accessibility in the Digital Age**
Dr. Alina von Davier, Chiel of Assessment at Duolingo.
Plenary Room 10

1:00pm-2:00pm **Graduate Student Committee Meeting & Lunch**
Conference Hall

2:00pm-3:15pm Session 2.1 **TOPIC Innovations in test development**
Room 1
Chair: Ricardo Rosas *(P. Universidad Católica de Chile)*

Interactions Between Big Five and Dark Side Traits Utilising Factor Analysis of Personality Assessment Data
Mikael Nederström *(Psycon/Finland)*

Playful testing of executive functions with Yellow-Red: Tablet-based battery for children between 6 and 12 available in 8 different languages
Ricardo Rosas *(P. Universidad Católica de Chile)*

Distractor Analysis of Eye Movements for Multiple-Choice Questions
Ergun Cihat Corbaci *(Dr. / Turkiye)*

2:00pm-3:15pm Session 2.2 **SYMPOSIUM Revising and enhancing the EFPA Test Review Model with lessons learned in practice**
Topic: International assessment
Room 2
Chair: Nigel Evans *(NEC)*

Implementation of the Test Review Model in Spain: Impact, Improvements and Challenges (Translation of tests, psychological assessment instruments and survey questionnaire)
Ana Hernández *(University of Valencia)*

Cross-Cultural Testing – recent observations from the British Psychological Society Test Review Process (Validity and fairness in cross-cultural testing, psychological assessment and survey research)
Charlie Eyre *(Consultant Editor, British Psychological Society Test Reviews, member of BPS Committee for Test Standards. Director, Workspheres Ltd.)*

Two applications of the EFPA Test Review Model in Norway (Translation of tests, psychological assessment instruments and survey questionnaire)
Siri S. Helland *(RBUP, Pilar, Norway)*

The revision of the EFPA Board of Assessment Test Review Model: the last hurdles on the way to a necessary and thorough update (International assessment)
Schittekatte Mark *(Ghent University, Belgium)*

2:00pm-3:15pm Session 2.4 **SYMPOSIUM Current Trends and Best Practices in Cross-Lingual Assessment**
Topic: Translation of tests, psychological assessment instruments and survey questionnaire
Room 4
Chair: Louise Badham (International Baccalaureate & University of Oxford / UK)

Maintaining Rigor and Comparability Across Different Language Versions of Language and Literature Assessments
Louise Badham *(International Baccalaureate & University of Oxford / UK)*

Developing and Evaluating Exams for Use Across Multiple Languages: The Successive (Adaptation) Model
Maria Elena Oliveri *(Buros Center for Testing)*

Developing, Facilitating, and Evaluating Validity and Comparability in Multi-Language Assessment Programs: A Review of the Literature
Stephen Sireci *(University of Massachusetts Amherst, USA)*

Best Practices in Adapting Test Across Languages: Lessons from cApStAn's Twenty-Five Years of Trial and Error
Steve Y. F. Dept *(cApStAn)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

2:00pm-3:15pm Session 2.5 **TOPIC** Psychometric modeling
Room 5
Chair: Miguel A. Sorrel *(Autonomous University of Madrid)*

Advances in testing MRI factorial models
Agustin Martinez-Molina *(Universidad Autónoma de Madrid)*

Modelling the relationship between problem-solving solution attainment and strategies using task-general process data with IRTrees
Huseyin Yildiz *(Australian Council for Educational Research/Australia)*

Monk: Developing a vertical scale based on IRT to assess Math competence in primary education in the Spanish education system
Miguel A. Sorrel *(Autonomous University of Madrid)*

Dimensionality Assessment in Forced Choice Questionnaires: First Steps Towards an Exploratory Framework
Diego F. Graña *(Universidad Autónoma de Madrid)*

2:00pm-3:15pm Session 2.6 **TOPIC** Artificial Intelligence in testing, psychological assessment and survey research
Room 6
Chair: Nancy Tippins *(The Nancy T. Tippins Group, LLC USA)*

Preserving the dignity and the privacy of candidates in AI-based assessments
Nancy Tippins *(The Nancy T. Tippins Group, LLC USA)*

From Job Descriptions to Facets: a Computational Technique for Effective Predictive Modeling
Tales Marra *(AssessFirst)*

Assessment Vulnerability to Cheating Using Large Language Models
Richard Landers *(University of Minnesota)*

2:00pm-3:15pm Session 2.7 **TOPIC** Testing equivalence by psychometrics methods
Room 7
Chair: Joshua Chiroma Gandi *(Department of Psychology)*

Rethinking Psychometric Properties in Developing Instruments for Optimal Assessment
Joshua Chiroma Gandi *(Department of Psychology)*

Investigating re-sitting effect on item difficulty in a medical selection test
Luc Le *(Australian Council for Educational Research)*

Fostering Fairness: Assessing measurement equivalence of MMIs
Mustafa Asil *(Bond University)*

Studying invariance with multiple groups in large datasets: A comparison of bayesian SEM and alignment method
Oscar Lecuona *(Universidad Complutense de Madrid)*

2:00pm-3:15pm Session 2.8 **TOPIC** Quantitative, qualitative, and mixed validation methods
Room 8
Chair: Anita M. Hubley *(University of British Columbia)*

A Conceptual Framework for Assessment Experience in Educational Testing
Fernando Mena Serrano *(University of Massachusetts Amherst)*

The effect of starting with easy items on SEM in a CAT
Serkan Arikan *(Bogazici University)*

Utilizing experts' judgments to determine cut-off scores of a test with little data
Julien Mouchnino *(Le français des affaires (CCI Paris Ile-de-France))*

Recent Trends, Emerging Controversies, and Consequences for Response Processes Validation: Lessons Learned
Anita M. Hubley *(University of British Columbia)*

Examining undergraduate students' programming process through cognitive interviews and keystrokes
Min Li *(University of Washington)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

2:00pm-3:15pm Session 2.9 **SYMPOSIUM** The assessment of cognitive abilities: critical approaches to validity and reliability
Topic: Validity theory in testing, psychological assessment and survey research
Room 9
Chair: Vanessa Torres van Grinsven *(Open Universiteit and University of Cologne)*

The Response Process in Standardized Cognitive Ability Tests and Validity
Vanessa Torres van Grinsven *(Open Universiteit and University of Cologne)*

The Test Tested
Anna M.T. Bosman *(Radboud University)*

Intelligence and the Individual
Anouk van Hoogdalem *(Radboud University)*

Dynamic assessment; An alternative to static testing?
Maartje Radstaake *(Radboud University)*

3:25pm-4:40pm Session 3.1 **SYMPOSIUM** Implementing a New Model for Measuring Clinical Judgment in Nursing: The Next Generation NCLEX
Topic: Innovations in test development
Room 1
Chair: Joe Betts *(NCSBN)*

The Best Seats in the House - A Viewpoint on the Development of the Next Generation Nursing Examinations (Validity theory in testing, psychological assessment and survey research)
April Zenisky *(University of Massachusetts Amherst)*

Developing and Testing Scoring and Response Models for Clinical Judgment Case Studies (Psychometric modeling)
Joe Betts *(NCSBN)*

Implementing a Polytomous CAT with Evolving Case Studies (Psychometric modeling)
Joe Betts *(NCSBN)*

Development of a Conceptual, Task, and Assessment Model for Measuring Clinical Judgment (Psychometric modeling)
William Muntean *(NCSBN)*

3:25pm-4:40pm Session 3.2 **SYMPOSIUM** Insights on changes in test use across Europe from the Work of the EFPA Board of Assessment
Topic: International assessment
Room 2
Chair: Nigel Evans *(NEC)*

How can we contribute to improving the quality of psychometric practices in the field of guidance and work: attempts, actions undertaken and reflection underway in France (Validity theory in testing, psychological assessment and survey research)
Even LOARER *(Cnam - Inetop, France)*

Online administration of tests of Italian psychologists: attitude and behaviours at the time of COVID-19 within an European Federation of Psychologists' Survey (Validity theory in testing, psychological assessment and survey research)
Adriana Lis *(Università di Padova)*

Challenges and Insights to Psychological Testing within Forensic Contexts in the UK (Validity and fairness in cross-cultural testing, psychological assessment and survey research)
Glenda Liell *(CTS, British Psychological Society)*

Revitalizing EFPA EuroTest Standards: Possible Strategies for Integrating Test User Standards into the EFPA Information Strategy (International assessment)
Urszula Brzezinska *(Pracownia Testow Psychologicznych PTP)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

3:25pm-4:40pm Session 3.3 **SYMPOSIUM / PANEL** Organized Discussion: Rethinking Conceptions of Fairness and Equity in Assessment
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research
Room 3
Chair: Randy Bennett *(ETS)*
Panelists: Randy Bennett *(ETS)*
Maria A. Ruiz-Primo *(Stanford University)*
Jennifer Randall *(University of Michigan)*
Ye Tong *(National Board of Medical Examiners)*

3:25pm-4:40pm Session 3.4 **SYMPOSIUM / ROUND TABLE** Adapting evaluation instruments to minority languages: Obstacles and alternatives
Topic: Translation of tests, psychological assessment instruments and survey questionnaire
Room 4
Chairs: Nekane Balluerka *(University of the Basque Country).* Jone Aliri *(University of the Basque Country) y* Arantxa Gorostiaga *(University of the Basque Country)*
Panelists: Nekane Balluerka *(University of the Basque Country)*
Jone Aliri *(University of the Basque Country)*
Arantxa Gorostiaga *(University of the Basque Country)*
Fréderique Vallar *(Pearson Clinical & Talent Assessment España)*
Pablo Santamaria *(Hogrefe TEA Ediciones)*
Maria E. Oliveri *(University of Nebraska, Lincoln)*
Mirko Antino *(The Spanish Journal of Psychology)*

3:25pm-4:40pm Session 3.5 **SYMPOSIUM** Computerized Adaptive Assessment of Cognitive and Non-Cognitive Competencies
Topic: Translation of tests, psychological assessment instruments and survey questionnaire
Room 5
Chair: Milagrosa Sánchez-Martín *(Universidad Loyola Andalucía)*, María Dolores Nieto-Cañaveras *(Nebrija University)*

Development of a Forced-Choice Item Bank for Adaptive Personality Assessment: A Pilot Study
Francisco J. Abad *(Universidad Autónoma de Madrid/España)*

Multicultural Adaptive Assessment of Teamwork Competencies
María Dolores Nieto-Cañaveras *(Nebrija University/España)*

Development and calibration of a new item pool to measure logical reasoning ability in undergraduate students
Milagrosa Sánchez-Martín *(Universidad Loyola Andalucía)*

Spelling competence in incoming university students: Creation of item pool based on current uses of Spanish spelling
Juan F. Luesia *(Universidad Loyola Andalucía)*

3:25pm-4:40pm Session 3.6 **TOPIC** Artificial Intelligence in testing, psychological assessment and survey research
Room 6
Chair: Denis Federiakin *(Johannes Gutenberg University of Mainz)*

Development and Validation of Automated Video Interview Competency Assessments in Spanish
Gema Ruiz de Huydobro *(HireVue)*

Parameterizing Linear Logistic Test Model as a Neural Network
Denis Federiakin *(Johannes Gutenberg University of Mainz)*

Machine Learning Unveils Factors Influencing Students' Math Performance Globally: Insights from PISA 2018
Liu Liu *(University of Washington)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

Differences in Perceptions of Artificial Intelligence (AI) - powered Assessments and Impact on Test Performance
Justine Chalifour *(HireVue)*

Predicting Missing Responses with Process Data in Large-Scale Computational Thinking Assessment
Qiwei He *(Georgetown University)*

3:25pm-4:40pm Session 3.8 **TOPIC** Quantitative, qualitative, and mixed validation methods
Room 8
Chair: Jon Twing *(University of Sydney)*

An Application of Differential Item Functioning to an Adolescent Assessment Adapted for Adult Learners
John Sabatini *(University of Memphis)*

Disentangling the Interplay of Emotional Intelligence, Personality Attributes, and MMIs in medical student selection
Mustafa Asil *(Bond University)*

An Application of G-Theory and Many Faceted Rasch Measurement in Performance Assessment
Jon Twing *(Unversity of Sydney)*

Developing the Facilitators and Obstacles of Recovery Scale (FOR-S) using the Delphi method: Insights from mental health professionals and service users
Georgina Guilera *(University of Barcelona, Spain)*

3:25pm-4:40pm Session 3.9 **SYMPOSIUM** College Admission in Chile: Inequity, Social Unrest, and More Testing
Topic: Quantitative, qualitative, and mixed validation methods
Room 9
Chair: Sergio Araneda *(Caveon Test Security)*

Standardized Testing and Social Equity: An Evaluation of Recent Changes in Chile's University Admissions (Identifying biases by qualitative or quantitative methods)
David Torres Irribarra *(Pontificia Universidad Católica de Chile)*

New Insights for Assessing the Predictive Capacity of Selection Tests in a Heterogeneous University System (Psychometric modeling)
Eduardo Alarcón-Bustamante *(Pontificia Universidad Católica de Chile)*

Studying Examinees' Experiences Shared on Tik Tok about Standardized Testing and College Admission in Chile (Quantitative, qualitative, and mixed validation methods)
Xaviera Gonzalez-Wegener *(Keele University)*

Change is never easy: the case of Chilean College admission system and their new battery of standardized tests PAES (Quantitative, qualitative, and mixed validation methods)
Fernanda Gandara *(Room to Read)*

4:40pm-5:30pm **Coffee break**

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

4:40pm-5:30pm **Poster session 2**

Room: Exhibition Hall

TOPIC Translation of tests, psychological assessment instruments and survey questionnaire

1 Validation of the Maternal Postpartum Stress Scale (MPSS) in Spanish population: analysis of the internal structure.
Adrian Ruiz-Perete, Sergio Martinez-Vazquez, Alejandro de la Torre-Luque, Rafael Caparros-Gonzalez

2 Evidence of content validity: Experts judgment in an Externalizing Problem Behavior Scale in Adults.
Lidia Torres Rosado, Cinta Mancheño Velasco, Andrea Blanc Molina, Manuel Sánchez García

TOPIC Quantitative, qualitative, and mixed validation methods

3 Are Open-Ended Demographic and Non-Demographic Items Useful in Evaluating Data Quality? Examining Responses from Adults Recruited via Amazon's MTurk.
Alexis D. Webster, Anita M. Hubley

4 Response Processes Validity Evidence Using Cognitive Interviews: Clear Reporting Necessary to Ensure Good Measurement Practice.
Sophie Ma Zhu, Amanda Rose Dumoulin, Anita M. Hubley

5 Identifying Imbalances and Gaps in Psychometric Evidence: A Reliability and Validation Synthesis of the Rosenberg Self-Esteem Scale.
Sophie Ma Zhu, Robert J. Ruddell, Anita M. Hubley

TOPIC Testing equivalence by psychometrics methods

6 Examining generational test score changes in Spatial and Word Analogy performance: Insights from Austrian conscript data.
Alina Bugelnig, Maria Gruber, Alexander Birner, Jakob Pietschnig

7 The English Version of the Satisfaction With Life Scale (SWLS) for Asian International Students in the United States: A Cross-Cultural Study.
Giusy Danila Valenti, Palmira Faraci

8 Factorial equivalence of the core cognitive abilities of the WAIS-IV across the US and UK.
Hannah Cruickshank Campbell, Christopher J. Wilson, Abigail Batty, Stephen C. Bowden

9 Comparing Item Parameters and Scales of Adolescent Assessment Adapted for Adult Learners.
John Sabatini

10 Modelling indicator-specific effects in longitudinal invariance: the case of the Revised-University of California at Los Angeles Loneliness scale.
Laura Galiana, Irene Fernández, Sara Martínez-Gregorio, Adrián García-Molla, José M. Tomás

11 Approximate invariance of ARDES measures in 7 countries using Alignment Analysis: Argentina, Australia, Brazil, China, Spain, UK, and USA.
Cándida Castro, Pablo Doncel, Rubén D. Ledesma, Silvana A. Montes, D. Daniela Barragan, Oscar Oviedo-Trespalacios, Alessandra Bianchi, Natalia Kauer, Weina Qu, Jose-Luis Padilla

12 A Psychometric Analysis of the CES-D 10 Using the South African National Income Dynamics Study (NIDS) Panel Data.
Richard Fletcher, Hermione Clayton, Natasha Bradley, Hayley Webb

TOPIC Psychometric modeling

13 An empirical comparison of IRT-based and Generalizability Theory-based approaches for estimating Conditional Standard Errors of Measurement in personality testing.
Gempp René

14 Moderated Nonlinear Factor Analysis for Measurement Invariance.
Joseph Kush

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

15 Matching student's vocational interests to study environments: Model-based mapping of university faculties in Holland's circumplex RIASEC space.
Lisa Bailey, Gideon de Bruin, Brandon Morgan

16 Rasch analysis of the 21 item Depression, Anxiety and Stress Scale in a mild traumatic brain injury sample.
Richard Siegert, Josh Faulkner, Deborah Snell

17 Measuring How Individuals Mentally Relate Science to Religion.
Rizqy Amelia Zein, Mario Gollwitzer

18 Character Strengths and Well-being: Does IRT Model Choice Influence the Shape of the Relationship?
Susanna Goosen, Gideon de Bruin

TOPIC Artificial Intelligence in testing, psychological assessment and survey research

19 Gender differences in personality assessment with language indicators from semantic vector subspaces: An invariance study.
Álvaro López-Herranz, José Ángel Martínez-Huertas

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

5:30pm-6:30pm Session 4.1 **SYMPOSIUM** Embedded Standard Setting and Item Alignment in Practice
Topic: Innovations in test development
Room 1
Chair: Ellen Forte *(edCount, LLC/USA)*

Embedded Standard Setting: Theory and Practice
Daniel Lewis *(Creative Measurement Solutions LLC/USA)*

Principled Alignment in Support of Validity and Coherence
Ellen Forte *(edCount, LLC)*

5:30pm-6:30pm Session 4.2 **SYMPOSIUM** Raising the Bar: Unveiling the New Quality Standards for PISA
Topic: International assessment
Room 2
Chair: Javier Suárez-Álvarez *(University of Massachusetts Amherst)*

Purpose and Scope of the New PISA Quality Standards
Ava Guez *(OECD)*

Using the new PISA Quality Standards to strengthen the development of PISA's innovative domain assessment
Mario Piacentini *(OECD)*

The Role of Fairness, Validity, Comparability, and Reliability in the New PISA Quality Standards
Javier Suárez-Álvarez *(University of Massachusetts Amherst)*

5:30pm-6:30pm Session 4.4 **TOPIC** Translation of tests, psychological assessment instruments and survey questionnaire
Room 4
Chair: Valeriia Manina *(HSE University)*

The Rasch/Guttman Scenario Approach to Developing the Scale of Ethical Behaviour in Organisations
Valeriia Manina *(HSE University)*

Can we predict Careless Responding? The role of sociodemographic variables
Clara Cuevas-Ureña *(University of Valencia)*

Succeeding in a matrix-reasoning task: effects of cognitive strategy and test characteristics
Natalie Badstuber *(Paris Lodron University of Salzburg)*

5:30pm-6:30pm Session 4.5 **TOPIC** Psychometric modeling
Room 5
Chair: Jorge González *(Pontificia Universidad Católica de Chile)*

More equitable and fairer measures of safe environment at schools in Chile
Jorge González *(Pontificia Universidad Católica de Chile)*

A comparison between Rasch Equating Method and Delta Scoring Equating Method using the Saudi National Assessment for Learning Outcomes
Ahmed Haddadi *(Education and Training Evaluation Commission)*

What can go wrong in a large-scale educational evaluation? Insights and recommendations from an educational assessment in Mexico
Scarlett Escudero *(Facultad de Psicología, Universidad Autónoma de Madrid/Spain)*

5:30pm-6:30pm Session 4.6 **SYMPOSIUM** Ethical issues in assessment arising from the use of AI
Topic: Artificial Intelligence in testing, psychological assessment and survey research
Room 6
Chair: Wayne Camara *(LSAC)*

Test taker data: Ownership and control of use of data from AI assessments
Wayne Camara *(LSAC/USA)*

Security, safety, and accountability in AI-based assessment
John Weiner *(Lifelong Learner USA)*

Fairness, transparency & explainability of AI-based assessment
Dragos Iliescu *(University of Bucharest Hungary)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

5:30pm-6:30pm Session 4.7 **TOPIC** Computational developments for social science research in cross-cultural testing
Room 7
Chair: Matthias von Davier *(TIMSS and PIRLS International Study Center)*

Advancing Education Assessment through NLP and AI: A Comprehensive Approach to TIMSS and PIRLS
Matthias von Davier *(TIMSS and PIRLS International Study Center)*

Automatic math item generator - Auto.Math: Bridging the gap between tradition and AI?
Steve Bernard *(University of Luxembourg)*

Multiple Imputation of missing values for randomized controlled trials: A step-by-step tutorial using mice
Oscar Lecuona *(Universidad Complutense de Madrid)*

5:30pm-6:30pm Session 4.8 **SYMPOSIUM** Testing the Periodic Table of Personality Methodology
Topic: Validity theory in testing, psychological assessment and survey research
Room 8
Chair: Rainer Kurz *(HUCAMA Analytics Ltd)*

Mapping Lumina Spark and Emotion Qualities to the TDA Periodic Table of Personality
Stewart Desson *(Lumina Learning)*

Personality Factors on the PF16 Periodic Table of Personality
Rainer Kurz *(HUCAMA Analytics Ltd)*

Normative and Ipsatised Great 8 Success Factors on the PF16 Periodic Table of Personality
Michele Guarini *(HUCAMA Group)*

5:30pm-6:30pm Session 4.9 **SYMPOSIUM** Alternative Approaches to Maximising Test Fairness
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research
Room 9
Chair: Jake Smith *(Saville Assessment Ltd, UK)*

Stereotypes and Testing Fairness
Sarah Chan *(Saville Assessment Ltd, UK)*

Fairer Assessment through Reducing Cognitive Load
Jake Smith *(Saville Assessment Ltd, UK)*

Implementing Fairer Single-stage Online Screening
Rab MacIver *(Saville Assessment Ltd, UK)*

7:15pm-10:00pm **Alhambra tour**
Meeting point: Granada Conference Palace

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

THURSDAY JULY 4th

8.30am-9.45am Session 5.1 **SYMPOSIUM** Continuous Norming: Recent Advancements in Research and Application – Part A
Topic: Innovations in test development
Room 1
Chair: Jan-Philipp Freudenstein *(Hogrefe Publishing Group)*

Where are we and Where to go? – A systematic Review and Real Data Example of Continuous Norming (Psychometric modeling)
Julian Urban *(Trier University)*

Comparing different Continuous Norming models in the creation of Rey Complex Figure Test (RCFT) norms (Psychometric modeling)
Yaiza Puig Navarro *(Hogrefe TEA Ediciones)*

The GRoNC-Checklist: Guidelines for Reporting on Norm-referenced and Criterion-referenced scores (Translation of tests, psychological assessment instruments and survey questionnaire)
Marieke Timmerman *(Psychometrics and Statistics, University of Groningen, the Netherlands)*

Zero-Inflated Beta-Binomial Distributions for Regression-Based Norming of Test Data with Floor and Ceiling Effects (Innovations in test development)
Jan-Philipp Freudenstein *(Hogrefe Publishing Group)*

8.30am-9.45am Session 5.2 **TOPIC** International assessment
Room 2
Chair: Marcus Henning *(University of Auckland)*

Understanding higher education students' ethical learning practices
Marcus Henning *(University of Auckland)*

Adopting ITC Guidelines for cross-cultural assessment: Translation and validation of an ADHD measure
Braden Hansma *(MHS)*

The Instructional Sensitivity of Constructed-Response Items in International Large-Scale Assessments
Anne Traynor (*Purdue University/United States)*

Do students respond inconsistently on mixed-worded scales in the PISA 2022 questionnaire?
Michalis Michaelides *(University of Cyprus)*

Integrating Intelligent Tutoring Systems with Active Learning to Enhance Testing Systems
Yoon Soo Park *(University of Illinois College of Medicine)*

8.30am-9.45am Session 5.3 **SYMPOSIUM** Comparability challenges in national and international summative assessment
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research
Room 3
Chair: Alejandro Veas *(University of Murcia/Spain)*

Application of the Rasch model to the Spanish university entrance examinations
Elena Govorova *(2E Estudios & Evaluaciones)*

Evaluating the cross-language comparability of PIRLS in South Africa
Heather Kayton *(University of Oxford)*

University entrance examinations and PISA: Analysis of curriculum and competences from the Spanish context (International assessment)
Jennifer Pérez-Sánchez *(University of Salamanca)*

Inter-subject comparability: small entry subjects and minority languages in the International Baccalaureate Diploma Programme (International assessment)
Louise Badham *(International Baccalaureate / University of Oxford)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

8.30am-9.45am Session 5.4 **SYMPOSIUM** Test translation and adaptation – survey methodology meets testing
Topic: Translation of tests, psychological assessment instruments and survey questionnaire
Room 4
Chair: Dorothée Behr *(GESIS - Leibniz Institute for the Social Sciences)*

Best practice in translation of survey instruments: 25 years of practice and research
Alisú Schoua-Glusberg *(Research Support Services Inc.)*

Advance translation - a method to enhance source questionnaire's translatability and cross-cultural portability
Brita Dorer *(GESIS-Leibniz Institute for the Social Sciences)*

Going beyond translation procedures – supporting comparability using (item-specific) translation guidance
Dorothée Behr *(GESIS - Leibniz Institute for the Social Sciences)*

Questionnaire translation in cross-cultural research: Translators, their background, and relevant competencies from the perspective of international research teams
Ulrike Efu Nkong *(GESIS Leibniz Institute for the Social Sciences)*

8.30am-9.45am Session 5.5 **SYMPOSIUM** Assessment in Organizational contexts
Topic: Validity theory in testing, psychological assessment and survey research
Room 5
Chair: María Dolores Nieto Cañaveras *(Nebrija University)*

Work Engagement revisited: A three item express scale
María Dolores Nieto-Cañaveras *(Nebrija University)*

Dimensionality of Organizational Climate
María Dolores Nieto-Cañaveras *(Nebrija University)*

Wellbeing at work can be measured
José Muñiz *(Nebrija University)*

Entrepreneurial Personality Assessment: The BEPE Battery
Álvaro Postigo *(University of Oviedo)*

8.30am-9.45am Session 5.6 **TOPIC** Artificial Intelligence in testing, psychological assessment and survey research
Room 6
Chair: Emeric Kubiak *(AssessFirst)*

Comparing Content Validity in Personality Assessment: Machine vs. Human-Authored Items using LLMs.
Simon Baron *(AssessFirst)*

Using ChatGPT for Semi-Automatic Generation of Items for Summative Assessment: a Case Study
Sergio Araneda *(Caveon Test Security)*

Unlocking Cognitive Patterns: A Comparative Exploration of Linear and Deep Dimensionality Reduction Approaches in clusterizing Student Cognitive Profiles
Matteo Orsoni *(University of Bologna)*

On the Use of Large Language Models to Generate Novel Collaborative Problem Solving Items
Yoav Bergner *(New York University)*

Using Natural Language Models to Assess Cognitive Complexity of Academic Standards
Kevin O'Rourke *(University of Massachusetts, Amherst)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

8.30am-9.45am Session 5.7 **TOPIC** Construct or concept equivalence
Room 7
Chair: Van Nguyen *(Australian Council for Educational Research)*

Evaluation of item time effect on a computer- based selection test
Van Nguyen *(Australian Council for Educational Research)*

Cross-cultural variability in lay perceptions of mental illness in Germany: Measurement invariance over cultural groups, gender, age, and education
Marie Kollek *(University of Hildesheim, Germany)*

Is It Worth to Use Constructed-Response Items in a Large-Scale Assessment Measuring Higher-Order Thinking?
Ozge Ersan *(Republic of Türkiye Ministry of National Education)*

How Cultural Cues Affect Bicultural Individuals' Personality Assessment Response Patterns: A Frame Switching Perspective
Patrick Lee *(Baruch College & The Graduate Center, CUNY)*

Using Q-matrices from CDM to inform the development of parallel test forms for the administration of diagnostic assessments in multiple languages
Sanet Steyn *(University of Cape Town)*

8.30am-9.45am Session 5.8 **SYMPOSIUM** Seeing Beyond Scores: Eye Tracking as a Gateway to Understanding Cognitive Processes and Attentional Dynamics of Test Takers
Topics: Quantitative, qualitative, and mixed validation methods /Translation of tests, psychological assessment instruments and survey questionnaire
Room 8
Chair: Marlit Annalena Lindner *(IWM, University of Tübingen)*

Investigating Response Models in Large-Scale Assessments: Refiguring Scale, Granularity and Diversity with Eye Tracking Studies
Bryan Maddox *(Assessment MicroAnalytics Ltd)*

Potential Contributions of Eye Movement Parameters in Measuring Ability via Multiple-Choice Items
Marlit Annalena Lindner *(IWM, University of Tübingen)*

Technology-Based Assessments (TBAs): Using eye movement data to understand test-taker attentional behaviour
Paula Lehane *(Dublin City University)*

The smell of paper or the shine of a screen? Students' reading comprehension, text processing, and attitudes when reading on paper and screen
Ragnhild Engdal Jensen *(University of Oslo)*

8.30am-9.45am Session 5.9 **TOPIC** Validity and fairness in cross-cultural testing, psychological assessment and survey research
Room 9
Chair: Yaseen Ally *(Nelson Mandela University)*

Can an adjustment index address the norming challenges in testing in South Africa? Investigating the use of an index to adjust scores
Yaseen Ally *(Nelson Mandela University)*

Decolonising Outcome Measurement with Kaupapa Māori Psychometrics: A systematic review and methodological quality appraisal of health and wellbeing measures for Māori
Richard Siegert *(Auckland University of Technology)*

Towards a Justice-oriented Approach to Assessment: Unpacking Religious Bias in English Language Assessments from Muslim Teachers and Students
Sheila Lallmamode *(AILLA Lab)*

Using Critical Quantitative Methodology and MIMIC Modeling for Justice-Oriented and Antiracist Measurement
Matthew Diemer *(University of Michigan)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

9.55am -10.55am	**INVITED PANEL** Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa **Plenary Room 10** **Chair:** Beatrice Rammstedt *(GESIS Deputy President, and Survey Design and Methodology Head of Department, Germany),* Ana Villar *(Survey Methodologist at Meta, UK),* Bruno D. Zumbo *(University of British Columbia Distinguished University Scholar and Professor, Canada)*
10.55am-11.45am	**Coffee break**

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

10.55am-11.45am **Poster session 3**

Room: Exhibition Hall

TOPIC Translation of tests, psychological assessment instruments and survey questionnaire

1 Exploratory Graph Analysis VS Exploratory Factor Analysis: A comparison of two methods for identifying the dimensional structure of the SHSS.
Alberto Sanchez-Fernandez-Quejo, Carla Perez-Guerra, Marco Innamorati, Alejandro de la Torre-Luque

2 Adaptation and Validation of the "Detail and Flexibility Questionnaire (DFlex)" in a Spanish Population.
Alicia Georghiades, Esteve Montasell, Ester Serrano, Beatriz Lanceta, Jordi Alabernia, Antoni Grau.

3 Spanish adaptation of the Hierarchical Taxonomy of Psychopathology-Self Report (HiTOP-SR): cultural issues across different Spanish-speaking countries. Ana María de la Rosa-Cáceres, Carmen Díaz-Batanero, Deisy Gonzalez-Zapata, Melissa Briones, Jennifer Callahan, Nazaret Fresno Cañada, Roman Kotov, Leonard Simms, B. Villalobos, Camilo J. Ruggero

4 Why is translating and analyzing the internal structure of scales not enough to certify their suitability in other cultures? An example with the Psychological Functioning Scale.
Ana Paula Noronha, Ligia Santis, Monique Guimarães, Ana Paula Cavallaro, Leila Couto

5 Focus on selection versus development for increased job performance. Four decades of research and a Monte Carlo simulation of individual differences and job attitudes.
Dragos Iliescu, Andreea Corbeanu, Andrei Ion

6 Psychometric Properties of the Spanish-SAM Motives Measure (S-SMM) among Young Adults who use cannabis.
Bella María González Ponce, Lucía Vélez-Pérez, Angelina Pilatti, José Carmona-Márquez

7 The Development of the SEL-90 Test.
Butucescu Andreea, Iliescu Dragos

8 Development of the Index of Intensity of Violence Against Women (IIVM) in Peru.
Carlos Renzo Rivera Calcina, Rodolfo J. Castro Salinas, Walter L. Arias Gallegos, Mitchell Clark

9 An Analisis of the Partner's Perceived Responsive Scale in Mexico.
Carolina Armenta Hurtarte, Pablo Tonathiu Salcedo Callado, María Bárbara Rivero Puente

10 Design and development of a Gamified Test to assess Critical Thinking in Personnel Selection Contexts.
Virginia Arranz, Sonia Rodríguez, Beatriz Lucía, David Aguado

11 Psychometric properties of the Screening Test for Reading and Spelling Difficulties for Lithuanian speakers in second grade.
Dovile Butkiene, Reda Gedutiene, Lauryna Rakickiene, Kestutis Dragunevicius, Grazina Gintiliene

12 The Zarit Burden Interview (ZBI-12): A Reliability Generalization Meta-analysis.
Elena Cejalvo Herraiz, Júlia Gisbert-Pérez, Laura Badenes-Ribera, Manuel Martí-Vilar

13 Psychometric properties and normative data of the Spanish S-UPPS-P Impulsive Behavior Scale in adolescents.
Esteve Montasell-Jordana, Eva Penelo, Laura Blanco-Hinojo, Anna Soler, Beatriz Lanceta, Alba Ollé, Jesús Pujol, Joan Deus

14 Critical consciousness in sport scale (CCSS): Development and psychometrics properties.
Evandro Peixoto, Martin Camiré

15 Functional Social Support assessment in Primary Health Care: a validation study of the DUKE-11-UNC in Spanish Primary health care users.
Irene Gómez Gómez , Juan Ángel Bellón, Isabel Benítez, Patrcia Moreno-Peral, Ana Clavería, Joan Llovera, José Ángel Maderuelo-Fernández, Rosa Magallón, Alvaro Sánchez Pérez, Bonabentura Bolibar, Emma Motrico

16 A psychometric analysis of the Emotion Regulation Questionnaire.
Jennifer Pérez-Sánchez, Ana R. Delgado, Gerardo Prieto

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

17 Transcultural adaptation of the REMICOM Test for the assessment of children's text comprehension monitoring.
Jesica Formoso, Nuria Calet Ruiz, Juan Pablo Barreyro, Bárbara Gottheil, Gracia Jiménez Fernández

18 Spanish validation of the Dragons of Inaction Psychological Barriers.
Jone Aliri, Laura Vozmediano, Laura Pasca, Olatz Goñi-Balentziaga

TOPIC Construct or concept equivalence

19 The Cross-Cultural Generalizability of Cognitive Ability Measures: A Systematic Literature Review.
Christopher Wilson, Stephen Bowden, Linda Byrne, Nicole Joshua, Wolfgang Marx, Lawrence Weiss

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

11.45am-1.00pm Session 6.1 **SYMPOSIUM** Continuous Norming: Recent Advancements in Research and Application – Part B
Topic: Innovations in test development
Room 1
Chair: Jan-Philipp Freudenstein (*Hogrefe Publishing Group*)

Adjusting for non-representativeness in continuous norming with Multilevel Regression and Poststratification (Psychometric modeling)
Klazien de Vries *(University of Groningen)*

Sample size calculation and optimal design for univariate and multivariate regression-based norming (Innovations in test development)
Francesco Innocenti *(Maastricht University)*

Item Response Theory Based Continuous Test Norming (Psychometric modeling)
Hannah Heister *(University of Groningen/Netherlands)*

More efficient continuous test norming by using prior norm information (Psychometric modeling)
Lieke Voncken *(Tilburg University)*

11.45am-1.00pm Session 6.2 **SYMPOSIUM** Debating Foundational Competencies in Educational Measurement: International Perspectives on an NCME Task Force Consensus
Topic: International assessment
Room 2
Chair: Andrew Ho *(Harvard University)*

Navigating the Foundational Competencies in Educational Measurement: Enhancing validity, validation, and fairness through comprehensive approaches (International Assessment)
Isabel Benitez *(University of Granada)*

On the future integration of foundational competencies frameworks in connected professions (International assessment)
Dragos Iliescu *(University of Bucharest)*

What is truly foundational: Continuing the conversation on the Foundational Competencies in Educational Measurement (International assessment)
Lisa Keller *(University of Massachusetts Amherst)*

Foundational Competencies in Educational Measurement: An NCME Task Force Consensus (International assessment)
Andrew Ho *(Harvard University)*

11.45am-2.00pm **Early Career Scholars Meeting**
Room 3

11.45am-1.00pm Session 6.4 **SYMPOSIUM** Exploring Human Dimensions - Love, Moral Dilemmas, Well-being, and Cultural Sympathy
Topic: Translation of tests, psychological assessment instruments and survey questionnaire
Room 4
Chair: Alejandra Dominguez Espinosa *(Universidad Iberoamericana)*

Validation of the Romantic Love Myths Scale of Bonilla-Algovia and Rivas-Rivero in a group of Mexican participants (Translation of tests, psychological assessment instruments and survey questionnaire)
Carolina Armenta Hurtarte *(Universidad Iberoamericana)*

Assessing tolerance to corruption through moral dilemmas: psychometric properties of a measurement (Translation of tests, psychological assessment instruments and survey questionnaire)
Alejandra del Carmen Domínguez-Espinosa *(Universidad Iberoamericana Ciudad de México)*

The measurement of subjective well-being in Mexico: A psychometric analysis of the ENBIARE 2021 (Validity theory in testing, psychological assessment and survey research)
José Luis López Silva *(UNAM)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

Proposal for Ethnopsychological Measurement to Estimate Agreeableness and Sympathy in Mexico (Validity and fairness in cross-cultural testing, psychological assessment and survey research)
Vanessa Edith Arellano Carranza *(Universidad Nacional Autónoma de México)*

11.45am-1.00pm Session 6.5 **TOPIC** Innovations in test development
Room 5
Chair: Nigel Evans *(NEC, UK)*

Advancing Precision and Validity with CAT and NLP
Alexandre Jaloto *(National Institute for Educational Studies and Research Anísio Teixeira (Inep))*

Privacy and Access to Data in Chile's National Admission Test: Balancing Students´ Rights and Institutional Needs
Agustin Barroilhet *(University of Chile)*

Measuring Communication and Interpersonal Skills of Pre-Service Teachers: An Experimental Study Using Frontal Alpha Asymmetry Indicators
Ahmet Haphap *(Gazi University/Turkey)*

Taking an Organimetric approach to Organisational Change
Nigel Evans *(NEC, UK)*

Qualitative procedures in developing a new emotion regulation measure for children from 4 to 6 years old
Denise Ruschel Bandeira *(Universidade Federal do Rio Grande do Sul)*

11.45am-1.00pm Session 6.6 **TOPICS** Translation of tests, psychological assessment instruments and survey questionnaire/ Validity and fairness in cross-cultural testing, psychological assessment and survey research
Room 6
Chair: Beatrice Rammstedt *(Deutschland)*

The association between changed translations and item functionality in ICCS 2022
Lauren Musu *(IEA)*

Going Global: 39 language versions of the BFI-2-XS
Beatrice Rammstedt *(Deutschland)*

SWIFT: Developing a Digital Psychometric Test for Swift Identification of Students at Risk for Learning Difficulties and Underachievement in Schools
Rebecca Good *(Education Elephant Ltd)*

Communication abilities in multilingual speakers with ADHD: insights from the Diagnostic interview for ADHD in adults
Maria Garraffa *(University of East Anglia and University of Oslo)*

Reliability and validity of the Test of Dyslexia for the Lithuanian speakers in second grade
Reda Gedutiene *(Klaipeda University, Lithuania)*

11.45am-1.00pm Session 6.7 **TOPIC** Identifying biases by qualitative or quantitative methods
Room 7
Chair: Ana Hernández *(University of Valencia. Spain)*

Reducing Evaluative Bias in Personality Assessment - Impact on Unboxing Neurodivergent Talent
Stewart Desson *(Lumina Learning)*

There's More Than Meets the Eye to Response Bias: Raising the Standards for Reporting the Integrity of Assessment Results
Carina Fiedeldey-Van Dijk *(Into Performance ULC)*

The Inexact Science of Fairness Panel Reviews: Can They Make Assessments More Culturally Relevant?
Jessica Jonson *(Buros Center for Testing - University of Nebraska-Lincoln)*

Careless responding: trait or state?
Inés Tomás Marco *(University of Valencia)*

Careless responding and DIF detection
Ana Hernández *(University of Valencia.)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

11.45am-1.00pm Session 6.8 **TOPICS** Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research
Room 8
Chair: Semih Topuz *(Baskent University)*

Fit for purpose or off the mark: PIRLS 2016 in South Africa
Heather Leigh Kayton *(University of Oxford)*

Integrating Validity Evidence for a Comprehensive Protocol to Assess Competencies in Incoming University Students
Juan F. Luesia *(Universidad Loyola Andalucía)*

MARKO-D a cross-cultural tool for early mathematics assessment
Victoria Espinoza *(Pontificia Universidad Católica de Chile)*

Exploring Careless Response Behaviours in Surveys: A Comparison of Different Identification Methods
Murat Doğan Şahin *(Anadolu University/Turkey)*

Does PISA Literacy Assessment Provide Fair Comparisons Across Participating Countries?
Semih Topuz *(Baskent University)*

11.45am-1.00pm Session 6.9 **SYMPOSIUM** Construct Validity of International WISC Versions: Informing Evidence Based Assessment
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research
Room 9
Chair: Gary Canivez *(Eastern Illinois University/USA)*

Construct Validity of the Canadian WISC–V with an Indigenous Sample: Hierarchical EFA and CFA
Gary Canivez *(Eastern Illinois University)*

Construct Validity of the Korean WISC–V: Hierarchical EFA and CFA
Gary Canivez *(Eastern Illinois University/USA)*

Construct Validity of the Australia/New Zealand WISC–V: Hierarchical EFA and CFA
Ryan McGill *(William & Mary)*

Construct Validity of the Brazilian WISC-IV: Hierarchical EFA and CFA
Solange Muglia Wechsler *(Pontificial Catholic University of Campinas-Brazil)*

1.00pm-2.00pm **Conference Lunch**

2.00pm-3.15pm Session 7.1 **SYMPOSIUM** Tech and Talent Synergy: Innovation for DEI, Individual, and Organizational Performance
Topic: Innovations in test development
Room 1
Chair: Maximilian Jansen *(Welliba)*, Richard Justenhoven *(Welliba)*

Perceptions of Image- and Questionnaire- Based Personality Measures Among Neurodivergent Adults
Franziska Leutner *(Goldsmiths)*

Advancing Fairness by Reducing Subgroup Differences with a New Logical Assessment
Mats Englund *(Fairsight/Sweden)*

Continuous Insights, Lasting Impact: The interaction of mindset and context factors in shaping Employee Experience
Maximilian Jansen *(Welliba/Germany)*

Job Satisfaction and Performance: What are the Great 8 Drivers of Job Success?
Michele Guarini *(Denmark)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

2.00pm-3.15pm Session 7.2 **ROUND TABLE** Challenges and Strategies: Measurement and Testing Associations in Focus
Topic: International assessment
Room 2
Chair: Nekane Balluerka *(University of the Basque Country)*
Panelists: Nigel Evans *(European Federation of Psychological Associations)*, Ana Hernández *(European Association of Methodology)*, Albert Sesé *(Spanish Association of Methodology)*, Stephen G. Sireci *(International Test Commission)*

2.00pm-3.15pm Session 7.3 **SYMPOSIUM** Bridging International Perspectives on Socioemotional Learning and Assessment
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research
Room 3
Chair: Javier Suárez-Álvarez *(University of Massachusetts Amherst)*

Assessing Socio-emotional Skills in Spain: Development and Validation of the Spanish Version of BESSI
Álvaro Postigo *(University of Oviedo)*

Validation and scaling of the OECD Survey on Social and Emotional Skills (SSES)
Elena Govorova *(2E Estudios & Evaluaciones)*

Assessing Social-emotional Skills using SENNA: Development, Psychometrics and Validity
Ricardo Primi *(Universidade São Francisco & EduLab21 Ayrton Senna Institute)*

2.00pm-3.15pm Session 7.4 **TOPIC** Translation of tests, psychological assessment instruments and survey questionnaire
Room 4
Chair: Therese Hopfenbeck *(University of Melbourne)*

Beyond Surveys: Multi-Modal Measures of 21st Century Skills Across Classrooms
Therese Hopfenbeck *(Assessment and Evaluation Research Centre, University of Melbourne, Australia)*

Evaluating the Quality of University Curriculum: A Theoretical Framework, Methodology, and Empirical Analysis
Wen Wen *(Tsinghua University)*

Utilizing the Ability to Identify Criteria (ATIC) to Select Personnel
Yolandi-Eloise Fontaine *(Stellenbsoch Univeristy)*

Ethical principles in Artificial Intelligence telerehabilitation for neurodevelopmental disorders: Development of a questionnaire for survey research
Aurora Castellani *(University of Perugia)*

2.00pm-3.15pm Session 7.5 **TOPIC** Translation of tests, psychological assessment instruments and survey questionnaire
Room 5
Chair: Camila Martínez *(Pontifical Catholic University of Chile)*

Integrating Transformer Models for Culturally Adaptive Trait Classification
Tales Marra *(AssessFirst)*

Executive functions in blind and deaf children: a Tablet-based assessment
Camila Martínez *(Pontifical Catholic University of Chile)*

Psychometric properties of Cattell's Fluid Intelligence Test (CFT 20-R) in Lithuanian sample of 8–17 year olds
Sigita Girdzijauskiene *(Vilnius University)*

Validation of the IDS-15 among Italian Adolescents
Adriana Lis *(University of Padova)*

What Does It Take To Make It: The Dark Side Of The Performing Arts
Melissa McMullan *(Edinburgh Napier University)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

2.00pm-3.15pm Session 7.6 **SYMPOSIUM** Generative AI and Large Language Models: Applications and Research Directions in Psychological Sciences
Topic: Artificial Intelligence in testing, psychological assessment and survey research
Room 6
Chair: Hudson Golino *(University of Virginia)*

Decoding Emotions: Facial Expression Recognition with Transformer Models using the transforEmotion Package in R (Artificial Intelligence in testing, psychological assessment and survey research)
Aleksandar Tomasevic *(Department of Sociology, University of Novi Sad)*

Monticello Simulations: How Generative AI change how we do simulations in quantitative psychology (Artificial Intelligence in testing, psychological assessment and survey research)
Hudson Golino *(University of Virginia)*

Assessing the Quality of AI-Generated Items: A Network Psychometric Approach (Artificial Intelligence in testing, psychological assessment and survey research)
Lara Russell-Lasalandra *(University of Virginia)*

Enhancing Student Evaluations of Teaching with Large Language Models: Insights from Active Learning Pedagogy (Artificial Intelligence in testing, psychological assessment and survey research)
Mariana Teles *(University of Virginia)*

2.00pm-3.15pm Session 7.7 **TOPIC** Translation of tests, psychological assessment instruments and survey questionnaire / Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Identifying biases by qualitative or quantitative methods
Room 7
Chair: Felipe Valentini *(University São Francisco, Brazil)*

Unveiling Cultural Biases: Exploring the Efficacy of Ipsative Methods in Recruitment Tools in Indonesia
R. Brahma Aditya *(Daya Dimensi Indonesia/Indonesia)*

Differential Prediction in an African Context: Analysing Personality and Sources of Bias in Selection
Andrew Morris *(JVR)*

The discrepancy between manifest response on a Likert-type scale and the most fixated response option as an indicator of social-desirable responding
Stanislav Ježek *(Faculty of Social Sciences, Masaryk University)*

A Methodological Approach to Evaluating and Selecting Overclaiming Items
Felipe Valentini *(University São Francisco, Brazil)*

Item Level Exploration of Applicant Faking Behaviour on Personality Measures: The importance of Cognitive Ability and Demographic Variables
Mollie Tatlow *(Thomas International/ UK)*

Public vs. Private School Dynamics: Insights into Anxiety and Depression Among Salvadoran Students.
Fernando Mena *(University of Massachusetts Amherst)*

2.00pm-3.15pm Session 7.8 **TOPIC** Validity theory in testing, psychological assessment and survey research
Room 8
Chair: Joshua Chiroma Gandi *(Nigerian Defence Academy, Kaduna, Nigeria)*

Psychometric Parsimonious Parameterization for Evidential Accuracy and Precision
Joshua Chiroma Gandi *(Nigerian Defence Academy, Kaduna, Nigeria)*

Advancing Linguistically and Culturally Fair and Community-Relevant Assessments
Pohai Kukea Shultz *(University of Hawaii/United States)*

Reduction of Faking Using a Forced-Choice Format: Is It Pancultural?
HyeSun Lee *(California State University Channel Islands)*

Factor Analysis under multimodal latent distributions: A simulation study
Oscar Lecuona *(Universidad Complutense de Madrid)*

Gender DIF and gender differences in civic outcomes over time
Yuan-Ling Liaw *(IEA Hamburg)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

2.00pm-3.15pm Session 7.9 **TOPIC** Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research
Room 9
Chair: Maria Elena Oliveri *(Buros Center for Testing)*

Development and Validation of an Integrative Employee Well-being Scale
Clara Y.W. To *(DIOP, The Hong Kong Psychological Society)*

Sociocultural Approaches to Assessing Work-Aligned Engineering Competencies
Maria Elena Oliveri *(Buros Center for Testing)*

Examining Methodologies for Contextual Information in Large-Scale Assessments
Priyanka Sharma *(India)*

Receptivity to Instructional Feedback: A Cross-cultural Validation Study
Carolina Lopera Oquendo *(City University of New York)*

Comparison of two methods to obtain the dimensionality of an instrument: factorial analysis Vs. Rasch analysis
Angélica Garzón Umerenkova *(Fundación Universitaria Konrad Lorenz - Facultad de Psicología (Bogotá-Colombia)*

3.25pm-4.40pm Session 8.1 **TOPIC** Innovations in test development
Room 1
Chair: Tatjana Kanonire *(HSE University)*

How modern psychometrics can change intelligence testing in children and adolescents
Tatjana Kanonire *(HSE University)*

Unlocking Natural Language Processing: Predicting Item Difficulty in the Brazilian High School Exam (Enem) Without Pre-Testing
Alexandre Jaloto *(National Institute for Educational Studies and Research Anísio Teixeira (Inep))*

Automatic Item Generation for Large-scale Assessment Instruments: A Mexican Perspective
Citlalli Sanchez-Alvarez *(Universidad Autónoma de Baja California/Mexico)*

3.25pm-4.40pm Session 8.2 **SYMPOSIUM** Personality and Potential Across Languages and Cultures
Topic: International assessment
Room 2
Chair: Lauren Jeffery-Smith *(Saville Assessment, United Kingdom)*

Assessing Leadership Potential Across Different Cultures (International assessment)
Lauren Jeffery-Smith *(Saville Assessment, United Kingdom)*

Multi-lingual Development of the Great 8 Success Factors (Translation of tests, psychological assessment instruments and survey questionnaire)
Rainer Kurz *(HUCAMA)*

Assessments and Machine vs Human Translation (Translation of tests, psychological assessment instruments and survey questionnaire)
Camille Stevenson *(Saville Assessment)*

3.25pm-4.40pm Session 8.3 **SYMPOSIUM** Test Validity 2.0: Mathematics, Quantum Information Theory, Movies, Music and Jokes
Topic: Validity theory in testing, psychological assessment and survey research
Room 3
Chair: Hudson Golino *(University of Virginia)*

Big Questions with Even Bigger Psychometricians: The Construct of Egomania Rageosis and The Theory of Everything
Bruno Zumbo *(University of British Columbia)*

Is there a natural law of validity or why Marlon Brando was so damn good as Don Corleone in The Godfather?
Hudson Golino *(University of Virginia)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

3.25pm-4.40pm Session 8.4 **TOPIC** International assessment
Room 4
Chair: Sharon Hague *(Pearson Assessments)*

TIMSS 2019 Equivalence Study: A Mixed-method Approach to Explore Assessment Mode Effects on Mathematics Performance in England
Grace Grima *(Pearson Education UK)*, **Liyuan Liu** *(Pearson Education UK)*

Estimating Missing Home Socioeconomic Status in PIRLS using Student and School Questionnaire Data: A MICE Simulation
João Marôco *(William James Centre for Research, ISPA - Instituto Universitário. Portugal)*

International Assessment Reform-A Case Study in Egypt
Sharon Hague *(Pearson Assessments)*

3.25pm-4.40pm Session 8.5 **TOPIC** Translation of tests, psychological assessment instruments and survey questionnaire
Room 5
Chair: Anita Obrycka *(1. World Hearing Center, Institute of Physiology and Pathology of Hearing, Kajetany/ Warsaw, Poland)*

Validation of AQoL-8D measures: a health-related quality of life questionnaire for adult patients referred for otolaryngology
Anita Obrycka *(1. World Hearing Center, Institute of Physiology and Pathology of Hearing, Kajetany/ Warsaw, Poland)*

Developing culturally sensitive and linguistically accurate versions of a music-based assessment for adults with disorders of consciousness
Wendy Magee *(Temple University, Philadelphia, USA)*

What is psychometric norm? Contemporary challenges for cultural adaptation of questionnaires for the assessment of mental disorders in adolescents
Joanna Stanczak *(Pracownia Testow Psychologicznych PTP)*

3.25pm-4.40pm Session 8.6 **SYMPOSIUM** Applications of Artificial Intelligence to Support Human and Automated Scoring of Constructed Responses
Topic: Artificial Intelligence in testing, psychological assessment and survey research
Room 6
Chair: Edward Wolfe *(Pearson)*

An Application of Data Augmentation to Automated Scoring Engine Training
Edward Wolfe *(Pearson)*

Curation of Validity Evidence for Applications of GPT4 in CR Scoring Systems
Jodi Casabianca-Marshall *(Educational Testing Service)*

Developing & Evaluating a Hybrid-Marking System to Combine AI & Human Scoring Models
Mark Brenchley *(Cambridge University Press & Assessment)*

3.25pm-4.40pm Session 8.7 **TOPIC** Construct or concept equivalence
Room 7
Chair: Velichko Fetvadjiev *(University of Amsterdam)*

The South African Personality Inventory (SAPI) across Cultures
Velichko Fetvadjiev *(University of Amsterdam)*

Assessment of collective efficacy and social cohesion in the context of policing
Miguel Inzunza *(Unit of Police Work/Umeå University/Sweden)*

Exploring Construct Equivalence of Items in Standardized Performance Testing: A Meta-Analytical Perspective on Response Formats
Sonja Breuer *(Paris Lodron University of Salzburg)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

3.25pm-4.40pm Session 8.9 **SYMPOSIUM** Equitable Selection into Initial Teacher Education Programs: The role of innovation, transparency and feedback
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research
Room 9
Chair: Therese Hopfenbeck *(University of Melbourne)*

Assessing the competencies and characteristics of prospective teachers: A fair selection process?
Laura Smith *(University of Melbourne)*

Technology and Contemporary Assessment: The challenges for Academic Integrity
Janet Clinton *(University of Melbourne)*

Developing evaluative thinking through sound assessment processes
Wayne Cotton *(University of Sydney)*

4.40pm-5.30pm **Coffee break**

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

4.40pm-5.30pm **Poster session 4**

Room: Exhibition Hall

TOPIC Validity and fairness in cross-cultural testing, psychological assessment and survey research

1 Are we really all that different? Examining Country and Gender Differential Item Functioning of the PROMIS Anxiety-8a.
Anita M. Hubley, Amanda Rose Dumoulin, Xuyan Tang, Sophie Ma Zhu

2 Hazard PredictionTest to assess drivers who suffered a stroke and healthy drivers.
Candida Castro, Daniel Salazar-Frías, Lucía Laffarga, Ana Szot, María Rodríguez Bailón

TOPIC Translation of tests, psychological assessment instruments and survey questionnaire

3 Psychometric Properties of the Spanish Version of the Cannabis Refusal Self-Efficacy Questionnaire (S-CRSEQ-13) among Young Adults who use cannabis.
Bella María González Ponce, Nehemías Romero-Pérez, Adrian J. Bravo, Fermín Fernández-Calderón

4 Assessing Teachers' Rational Number Knowledge: An Evaluation of Initial Validity Evidence.
Joanne Joo, Leanne Ketterlin Geller, Sarah Powell, Erica Lembke

5 Psychometric properties of the Smartphone Addiction Scale (SAS-SV) in adolescents in Peru.
Joel Figueroa-Quiñones

6 Analyzing Age-Based Measurement Invariance: A Study of the Basque Adaptation of the GPIUS-2 in Problematic Internet Use.
Jone Aliri, Olatz Goñi-Balentziaga, Nekane Balluerka, Arantxa Gorostiaga

7 Competitive Latent Structures for the Comic Style Markers: Developing a Psychometrically Sound Short Version Using Spanish and US American Samples.
Jorge Torres-Marín, Ginés Navarro-Carrillo, Mariela Bustos-Ortega, Sonja Heintz, Hugo Carretero-Dios

8 Video Game Dependency Scale: A Reliability Generalization Meta-analysis.
Júlia Gisbert-Pérez, Elena Cejalvo, Manuel Martí-Vilar, Laura Badenes-Ribera

9 Operational definitions and measurement of Externalizing: an integrative review and a new proposal.
Lidia Torres Rosado, Cinta Mancheño Velasco, Alberto Parrado González, Óscar Lozano Rojas

10 Voluntary Simplicity and Social Psychology: creation of a scale.
Luis Mundi López, Chiara Ambrosio, Eva Moreno-Bella, Josefa Ruiz Romero, Andrea Velandia Morales, Guillermo Willis Sánchez.

11 Validation of the Spanish version of the Emotional Style Questionnaire.
Maria Dolores Lopez-Martinez, Maria Dolores Hidalgo-Montesinos

12 Cross-Cultural Adaptation of the Perth Alexithymia Questionnaire (PAQ) for the Brazilian Context.
Maria Julia De Melo Amorim Venâncio, Cristiane Faiad de Moura

13 Optimization of a self-regulation for learning online scale using IRT information.
Mariel Fernanda Musso, Eduardo C. Cascallar

14 Psychometric properties of the EPPIP Scale in Health Sciences students from a Peruvian university.
Mercedes Merryl Jesus Peña, Antonella Alexandra Gallegos Arteaga, Rabbi Robinson Reyes Robles

15 The Kuwaiti-Arabic Trait Emotional Intelligence Questionnaire-Short Form: The Adaptation and Validation of the TEIQue-SF in Kuwait.
Nasser Hasan, Konstantinos Petrides

16 The Illness Management and Recovery Scale: Translation and validation study of the Spanish version.
Nuria Martín Ordiales, Mª Dolores Hidalgo Montesinos, Maite Barrios Cerrejon , Mª Pilar Martín Chaparro

17 The Screen for Cognitive Impairment in Psychiatry-Spanish in older adults: Preliminary analysis of its dimensionality and internal consistency.
Oscar Pino, Georgina Guilera, Vanessa Sanz, Maria Guallart, Emilio Rojo, Juana Gómez-Benito

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

18 Psychometric properties of ITQ and IPQ: Adaptations to the Spanish context to measure efficacy of Virtual Environments to generate emotional states.
Pablo Doncel, Miguel-Angel Muñoz, Maria Blasa Sanchez-Barrera, Pedro Garcia-Fernandez, Francisco Gómez, Jolanda Tromp, Daniel Salazar-Frías, Andreea Ionela Dinu, Candida Castro

19 Perceived utility of Q-matrices in the translation of diagnostic assessments: Translating an academic literacy test used for diagnostic purposes.
Sanet Steyn

5.30pm-6.30pm	**ITC General Meeting** **Room 10**
8.30pm	**Gala Dinner** (Carmen de los Mártires)

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

FRIDAY JULY 5th

8.30am-9.45am Session 9.1 **SYMPOSIUM** MatriKS: a New Computerized Raven-like Test for the Efficient Assessment of Fluid Intelligence
Topic: Innovations in test development
Room 1
Chair: Debora de Chiusole *(University of Padua)*

Convergent and divergent validity of MatriKS: A new tool to assess fluid intelligence
Alice Bacherini *(University of Perugia, Italy)*

Enhancing the assessment of fluid intelligence with MatriKS: Insights from knowledge space theory and multi-method analysis techniques
Debora de Chiusole *(University of Padua)*

Italian adaptation of the System Usability and Acceptance Model scale: application to MatriKS a new digital test for fluid intelligence assessment.
Matilde Spinoso *(Department of Psychology Renzo Canestrari, University of Bologna)*

Developmental trajectories of accuracy and type of errors in fluid intelligence assessment as detected by a new digital tool: MatriKS
Noemi Mazzoni *(University of Bologna)*

8.30am-9.45am Session 9.2 **SYMPOSIUM** Implementation of Teacher Performance Assessments Internationally
Topic: International assessment
Room 2
Chair: Jon Twing *(University of Sydney)*

Teacher Preparation and Performance Assessment
Mark Grant *(Australian Institute for Teaching and School Leadership, AITSL)*

Assessment for Graduate Teaching (AfGT)
Janet Clinton *(University of Melbourne)*

A TPA called the AfGT with AI & ML
Wayne Cotton *(University of Sydney)*

The Similarity and Differences in Teacher Performance Assessments Internationally
Jon Twing *(University of Sydney)*

8.30am-9.45am Session 9.3 **TOPIC** Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research
Room 3
Chair: Solange Muglia Wechsler *(Pontificial Catholic University of Campinas-Brazil)*

Psychological assessment in Latin American countries: Perspectives and Challenges
Solange Muglia Wechsler *(Pontificial Catholic University of Campinas-Brazil)*

Innovative Development and Validation of Tale Me More®, a Multicultural Professional Profiling Model: Integrating Theory and Practice
Céline Jouffray *(Talent Tale - France)*

Innovative Development and Validation of Tale Me More®, a Multicultural Professional Profiling Model: Integrating Theory and Practice
Wendy Magee *(Temple University Philadelphia USA)*

Examining the Predictive Validity of the Romanian Adaptation of the Conners-3 Short Form for ADHD Diagnosis
Serban Zanfirescu Zanfirescu *(University of Bucharest)*

Validity evidence of the Pornography Consumption Inventory: Relationship with subjective and objective sexual arousal measures
Oscar Cervilla Saez *(Mind, Brain and Behavior Research Center, CIMCYC)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

8.30am-9.45am Session 9.4 **TOPIC** Translation of tests, psychological assessment instruments and survey questionnaire
Room 4
Chair: Joshua McGrane *(The University of Melbourne)*

Evaluating an AI-based method for predicting language DIF in high-stakes, cross-language science assessments
Joshua McGrane *(The University of Melbourne)*

Asymmetrical Item Response Models
Jorge Bazán *(University of São Paulo)*

Cross-Cultural Validation: English version of the Problematic Use of Social Networking Sites Questionnaire
Covadonga González-Nuevo *(University of Burgos)*

Psychometric Investigation of Fairness: Role of Test Content Language and Student's Language
Jordan Southcott *(Multi-Health Systems)*

Beach Centre Family Quality of Life Scale: Urdu Translation, Adaptation, and Validation in the Context of Mental Illness in Pakistan
Rabia Khawar *(Department of Applied Psychology, Government College University Faisalabad, Pakistan)*

8.30am-9.45am Session 9.5 **SYMPOSIUM** Forced-Choice Measurement - Investigating Response Processes
Topic: Psychometric modeling
Room 5
Chair: Susanne Frick *(TU Dortmund University)*

Modelling 'intermittent faking' on forced-choice questionnaires
Anna Brown *(University of Kent)*

Trifactor Change Models for Likert and Force Choice Questionnaires
Nigel Guenole *(Goldsmiths, University of London)*

Careless Responding in Multidimensional Forced-Choice Questionnaires: What Does it Look Like and how can it be Detected?
Rebekka Kupffer *(University of Kaiserslautern-Landau)*

Using Process Data to Understand Response Processes Underlying Faking in Questionnaires
Susanne Frick *(TU Dortmund University)*

9.00am-9.45am Session 9.6 **TOPIC** Artificial Intelligence in testing, psychological assessment and survey research/ Innovations in test development
Room 6
Chair: Ken Clark *(Responsive Translation)*

Innovations in test development
Emeric Kubiak *(AssessFirst)*

Ambiguity in Educational Assessments: The Impact of LLM AI on Source Preprocessing in the Translation Workflow
Ken Clark *(Responsive Translation / United States of America)*

Establishing Validity Arguments in Automated Scoring Contexts: A Roadmap
Hillary Michaels *(Human Resources Research Organization)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

8.30am-9.45am Session 9.7 **TOPIC** Identifying biases by qualitative or quantitative methods/ Translation of tests, psychological assessment instruments and survey questionnaire
Room 7
Chair: Ronja Runge *(University of Hildesheim, Germany)*

Anchoring Vignettes: A Useful Tool to Measure and Correct for Cultural Bias in Parent Reports on Their Child's Mental Health?
Ronja Runge *(University of Hildesheim, Germany)*

Relationships between illegitimate tasks and employees' psychological distress: A Cross-level moderated mediating model
Yingwu Li *(Renmin Univerisity of China)*

Developing cutoff points to interpret impairment associated with depression and anxiety symptoms according to sex using the IDAS-II
Ana María de la Rosa Cáceres *(University of Huelva, Spain)*

Validation of depression anxiety stress scale (DASS 42) and the brief religious coping inventory (RCOPE) among Ghanaian population: Application of classical measurement and item response theories
Regina Mawusi Nugba *(University of Cape Coast)*

8.30am-9.45am Session 9.8 **TOPIC** Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research
Room 8
Chair: Jasmin Kalar *(Thomas International)*

Investigating the relative predictive validity of behaviour, personality and values upon work-related outcomes
Jasmin Kalar *(Thomas International)*

Linking Personality to PsyCap: Validation of a PsyCap-based Personality Assessment using CPAI-2
Clara Y.W. To *(The Hong Kong Psychological Society)*

Development and Validation of Chinese Pictorial Big Six Personality Inventory for Children (CPBSI-C)
Weiqi Mu *(CAS Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China)*

Presentation and validation of the self-capacity scale, adaptation in cross-cultural situations
Nicolas Drouin *(France)*

The Acceptance of Myths About Cyber-Sexual Violence Against Women in Spain and the United States: A Measure Invariance Study
Rocío Vizcaíno-Cuenca *(University of Granada)*

8.30am-9.45am Session 9.9 **TOPIC** Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research
Room 9
Chair: Nicola Taylor *(University of Johannesburg)*

The predictive value of the Five Factor Model across the lifespan
Nicola Taylor *(University of Johannesburg)*

Heterogeneity of scale items is biasing estimates of Omega consistency
Karl Schweizer *(Goethe University Frankfurt)*

Are Participants Fully Engaged During a Test Process? A Sequential Argument and Comparison with Different Methods
Murat Doğan ŞAHIN *(Anadolu University/Turkey)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

9.55am-10.55	**KEYNOTE Seeking Validity Evidence Related to Assessment Justice** Jennifer Randall *(Dunn Family Endowed Professor of Psychometrics and Test Development, University of Michigan, USA)* **Plenary Room 10**
10.55am-11.45am	**Coffee break**

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

10.55am-11.45am **Poster session 5**

TOPIC Validity and fairness in cross-cultural testing, psychological assessment and survey research

1 Humility and Modesty: Characteristic traits of the Mexican.
Bianca Pérez

2 Evaluating Sociocultural Influences on Student Interpretation in a Science Vocabulary Measure: The Utility of a Coding Rubric.
Jose Palma, Doris Baker, Holland Kowalkowski

3 Validation of the ICF Core Set for schizophrenia from the perspective of relatives: An international study.
Karina Campoverde, A.Chai Chuen, Juana Gómez-Benito, Maite Barrios, Georgina Guilera, Emilio Rojo

4 Cross-Cultural Assessment: Psychometric Requirements for a Cross-Cultural Norm of the TOP.
Kilian Hasselhorn, Jan-Philipp Freudenstein

5 Developing and obtaining validity and reliability evidence of the Multiple-Choice Test Quality Scale (ECET).
Paula Muñoz Teno.

6 Development and psychometric analysis of a new scale to measure adolescents' social and public commitment for environment.
Sofia Santisi, Caterina Primi, Angelo Panno, Luciano Romano, Maria Anna Donati

7 Harmonising the measurement of quality of life in SHARE across European regions.
Zaira Torres, Irene Fernández, Adrián García-Mollá, Amparo Oliver, José M. Tomás

TOPIC Validity theory in testing, psychological assessment and survey research

8 RPS-MAP (Route Planning Strategies in a Map) TEST: Sensitive assessment tool to identify cognitive-functional impairment related to route planning in Stroke drivers.
Lucía Laffarga, Ana Clara Szot, María Rodríguez-Bailón, Daniel A. Salazar-Frías, Pablo Doncel

9 Measurement invariance of the WISC-V in Chile: A contribution to fairness in psychological assessment.
Marcela Rodríguez-Cancino, Andrés Concha-Salgado

10 Comparison of Threshold Identification Methods for Response Time Effort across PISA Item Types: Evaluation Based on Validity Evidence.
Militsa Ivanova, Michalis Michaelides, Hanna Eklöf

11 Validation of the Pornography Consumption Inventory in the Spanish adult population.
Oscar Cervilla Saez, Ana Álvarez-Muelas, Laura E. Muñoz-García, Pablo Mangas, Gracia M. Sánchez-Pérez, Juan Carlos Sierra

12 Psychometric Properties of the Scientific Reasoning Scale: Application to the Italian Context.
Rossella Caliciuri, Margherita Lanz

TOPIC Psychometric modeling

13 A B-ESEM Model for the Impact of Event Scale-Revised (IES-R): New Conceptual and Methodological Perspectives on a Popular Cross-Cultural Measure for PTSD.
Giusy Danila Valenti, Palmira Faraci

TOPIC Translation of tests, psychological assessment instruments and survey questionnaire

14 Creation of an International Protocol for Assessing Tests, Scales, and Questionnaires (PETEYC).
Elena Govorova, Elena de la Guía, Gloria García-Moreno, Sara Díaz, Isabel Benítez

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

11.45pm-1.00pm Session 10.1 **TOPIC** Innovations in test development
Room 1
Chair: Anita Rintala-Rasmus *(Psycon / Finland)*

The Validation of a Multidimensional 360-Degree Assessment Tool and Its Relationships with Supervisor Engagement, Involvement, and Burnout Tendencies
Anita Rintala-Rasmus *(Psycon / Finland)*

The identification of gifted students through a large-scale educational test: an analysis of sociodemographic characteristics
Tatiana Nakano *(Pontifical Catholic University of Campinas)*

Reforming High-stakes, Low Volume Tests
Lei Yu *(U.S.)*

Building Culturally Sustaining Assessments to Support Adult Learners: From Co-design Studies to Assessment Development and Profile Reporting
Stephen G. Sireci *(University of Massachusetts Amherst)*

11.45pm-1.00pm Session 10.3 Scholars
Room 3
Chair: Felipe Valentini

Identifying psychological factors that improve mathematics achievement in Grade 9 pupils from Gauteng (Quantitative, qualitative, and mixed validation methods)
Pakeezah Rajab *(Senior Researcher at JVR Psychometrics and PhD candidate at University of Pretoria)*

Five-Factor Narcissism Inventories: Psychometric Properties of the BrazilianPortuguese Versions (Translation of tests, psychological assestment instruments and survey questionnaire)
Ariela R. Lima-Costa *(São Francisco University, Campinas, Brazil)*

Balancing test-taking experience and measurement efficiency in computerized adaptive testing: should easier adaptive tests be used? (Testing equivalence by psychometrics methods)
Hanif Akhtar *(ELTE Eötvös Loránd University, Hungary. University of Muhammadiyah Malang, Indonesia)*

11.45pm-1.00pm Session 10.4 **TOPIC** Translation of tests, psychological assessment instruments and survey questionnaire
Room 4
Chair: Jaime García-Fernández *(University of Oviedo)*

Spanish Adaptation of the Short-Dark Tetrad (SD4)
Jaime García-Fernández *(University of Oviedo)*

Diagnostic Adaptive Behavior Scale: Italian validation and standardization
Giulia Balboni *(University of Perugia/Italy)*

Validation of the Dyslexia Screeing Test-Junior (DST-J) in an Arabic-speaking context
Mahmoud Amer *(Sultan Qaboos University)*

Thinking and Creating Styles: Assessment in Portugal
Margarida Pocinho *(University of Madeira)*

Updates on the Transcultural Adaptation and Validation of the Diagnostic Adaptive Behavior Scale (DABS) for Brazil
Denise Ruschel Bandeira *(Universidade Federal do Rio Grande do Sul)*

11.45pm-1.00pm Session 10.5 **SYMPOSIUM** Forced-Choice Measurement - Challenges in Test Development
Topic: Psychometric modeling
Room 5
Chair: Markus Jansen *(University of Wuppertal)*

Developing a forced-choice personality questionnaire for the Austrian military (Psychometric modeling)
Christian Ludwig Becker *(Cupio OG – Psychologie für die Praxis)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

Construction and Validation of the HEXACO-MFC (Translation of tests, psychological assessment instruments and survey questionnaire)
Eunike Wetzel *(University of Kaiserslautern-Landau)*

Matching Items by Social Desirability Rankings to Improve the Faking Resistance of Multidimensional Forced Choice Questionnaires (Translation of tests, psychological assessment instruments and survey questionnaire)
Killisch Jan *(RPTU Kaiserslautern-Landau)*

Test and item design in forced-choice modeling with Thurstonian linked blocks (Innovations in test development)
Markus Jansen *(University of Wuppertal)*

11.45pm-1.00pm Session 10.7 **TOPIC** Testing equivalence by psychometrics methods
Room 7
Chair: Hannah Cruickshank Campbell *(Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, Australia)*

Measurement Invariance of the Wechsler Adult Intelligence Scale–Fourth edition across US and Spain Nationally Representative Samples
Hannah Cruickshank Campbell *(Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, Australia)*

Cross-National Generalizability of the WISC-V & CHC Broad Ability Constructs Across France, Spain, and the US
Christopher Wilson *(Pearson Clinical Assessment / Australia)*

Measuring well-being in school across all school ages
Tatjana Kanonire *(HSE University)*

Non-equivalence in PISA's ESCS index: no longer comparing apples with apples
Gavin Brown *(The University of Auckland/New Zealand)*

Use of Computerised Adaptive Testing in the Content of Student Evaluations of Teaching at Higher Education
Ilker Kalender *(Bilkent University, Faculty of Education)*

11.45pm-1.00pm Session 10.8 **TOPIC** Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research
Room 8
Chair: Kurt Geisinger *(Buros Center for Testing, University of Nebraska-Lincoln, USA)*

The length and verbal labels do not matter: The influence of various Likert-like response formats on scales' psychometric properties
Hynek Cígler *(Masaryk University, Czech Republic)*

Examining Validity Evidence of Constructs in Applied Linguistics: A Systematic Review
Angel Arias *(Carleton University / Canada)*

Applying Multidimensional IRT Models in Validating the Dimensionality of the Future Skills Exam
Fathima Jaffari *(Senior Measurement specialist, department of Tests and Measurement, National Center for Assessment, Education and Training Evaluation Commission (ETEC), Riyadh, Saudi Arabia)*

Generative item models as learning opportunity: a psychometric analysis of 25 item models in the mathematical domain of space & form
Philipp Sonnleitner *(University of Luxembourg/ Luxembourg)*

Testing Individuals with Disabilities for International University Admissions: Learning from the United States Experience
Kurt Geisinger *(Buros Center for Testing, University of Nebraska-Lincoln, USA)*

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

1.10pm-2.10pm	**SPECIAL SESSION: "Thomas Oakland Award and Closing Ceremony"**
	Plenary Room 10
	Presentation of the Thomas Oakland Award Winner. **Mª Dolores Hidalgo,** *University of Murcia, Spain. President of the Award Committee*
	Presentation by the Thomas Oakland Award Winner
	Presentation of the Best poster Award. **Albert Sesé,** *University of the Balearic Islands, Spain. Best Poster Awards Committee*
	Closing by a representative from the University of Granada.
	Closing by the ITC President-Elect. **Kadriye Ercikan,** *ETS, USA*
3.15pm-4.30pm	**ITC Council Meeting (Council Members Only)**

Note: *Only presenters are indicated in the program. All the authors will be specified at the book of abstracts*

ITC CONFERENCE

02 · 05 JULY 2024

WEDNESDAY 3 JULY
Session 1.1.
Topic: Innovations in test development

67. **Deployed Confidence: The Testing in Production**

Shreya Asthana
Red Hat

How do testers know that the feature they tested on stage is working perfectly in production as well? If something breaks in production, how will you know? The panic mode starts when your staging test results do not reflect current production behavior. Right? And you started doubting your testing skills when finally the user reported a bug to you. Have you tried testing in production? Yes you heard it right, "Testing in PRODUCTION". Once you start doing testing in production, you will see accuracy of test results, your tests will run faster due to elimination of bad data, and you will have higher confidence before releases. This can be accomplished through feature flagging, canary releases, and data cleanup. You will leave this talk with strategies to understand the steps to achieve production testing before making your feature live, and to shift your company's testing culture, so you can provide the best possible experience to the end users. This talk is beneficial because too many people think that testing should be done in staging but not in production and now this is high time to pull out people from their old mindset of testing into a new testing world. At the end of the day, we don't care if your features work in staging, we care if they work in production.

WEDNESDAY 3 JULY
Session 1.1.
Topic: Innovations in test development

90. Evaluating and Evolving Cognitive Assessments in the Age of Large Language Models

Emeric Kubiak, Simon Baron
AssessFirst

Objective. Large Language Models (LLMs) have gained interest regarding their ability to reason close to the human level. Research suggests that LLMs achieve state-of-the-art performance in quantitative reasoning (Lewkowycz & al. 2022) or that chain-of-thought prompting help LLMs perform better on different reasoning tasks (Wei & al., 2023). However, they perform poorly on multi-step problems (Creswell & al, 2022) and can't plan (Valmeekam & al., 2022). Still, we must recognize the impact LLMs have on current tools from psychology (Binz & Schulz, 2022), especially cognitive assessments, considered one of the main predictors of job performance (Schmidt & al, 2016). Our goal is then to (1) study how LLMs perform on MCQ cognitive assessment, (2) develop a new kind of assessment that LLMs can't answer. Study 1. We used a_test from an online platform, which includes analogy, verbal, quantitative, and abstract tasks (X = .59; α = .9). Items were entered as prompts into ChaGPT 4.0. It solved 79% of the items (30 out of 38) with poorer results on abstract reasoning. Study 2. It presents the construction of an assessment, which: (1) measure *g*, (2) is composed of logical reasoning tasks respondents must answer by combining material at their disposal, (3) is adaptive, (4) free from verbal information that is easily understandable by LLMs, (5) take 10 minutes to complete. This test has been developed by testing 400 items (N = 8,000). The final item bank is composed of 76 items. It shows good convergent validity (r = .73) with a progressive matrices test generated using IMak (= .84). Items were entered as image prompts into ChaGPT 4.0. It solved 0% of the items. Conclusion. Our work has (1) theoretical implications, by discussing the impact of LLMs in psychometrics, (2) practical implications, as this kind of assessment could help companies propose a more accurate hiring process and avoid AI-based faking.

WEDNESDAY 3 JULY
Session 1.1.
Topic: Innovations in test development

192. A Cross-Cultural Test explained by a Scientific Dialectical Discourse

Claudia Gusso
International Test Commission (ITC)

A cross-cultural testing procedure could be favorably carried out starting from the hypothesis that a synoptic model of the psycho-corporeal human scheme has been designed. In the analysis phase, a synoptic system used as a representation of the human model reduces the multiple dimensions and allows to identify and compare test data at a fundamental level. The research thesis explains how paths of divergence are oriented to respect cultural differentiation in a translation and adaptation process; these paths activate the production of a large quantity and quality of useful data conveyed towards the different containers suitable for accommodating the typical meanings of each culture. In the synthesis phase, experts declare that the model is anchored to the scientific principles of measurement in psychology and human physics: the method drives test users to interpret a person's characteristics, whose personality is a complex connection between psychic and physical aspects. We will analyze examples of translation and adaptation in settings of tri-cultural divergence through European (ITA), Asian (PRC), and American (USA) samples. The scientific dialectical discourse continues in the synchresis phase, with a debate on standard measurement in a cross-cultural context: considerations on the moral recognition of diversity will be presented, comparing the suitable elements chosen to define the universal characteristics of the human model as a metric reference standard. Why and how does this research catalyze applications for translation and cross-cultural adaptation of testing? Since its conception, the original Static Dynamic Personality Test (TPSD) has provided innovative psychometric measure solutions, which have been implemented to facilitate the cross-cultural procedure of test translation and adaptation. CLAUDIA GUSSO (0000-0003-0489-0963) – ORCID

WEDNESDAY 3 JULY
Session 1.1.
Topic: Innovations in test development

207. Does Practice Make Perfect? Unraveling the Influence of Practice Tests on Psychometric Assessment Accuracy and Equity

Karim Badr, Darrin Grelle
SHL

Many factors such as test anxiety, lack of familiarity with a test format, and a clear understanding by candidates of what the test is measuring can all influence the accuracy and validity of measurement of a given construct in a psychometric assessment. These factors can be mitigated through providing candidates the opportunity to take practice tests before sitting for the actual assessment (e.g., Yang et al., 2023). Cognitive ability tests used for personnel selection often involve items that candidates might not have had prior exposure to in formal education, so providing opportunities for the candidates to practice those items is warranted. We analyzed data collected in the last two years from over 100,000 candidates taking optional practice tests right before sitting their cognitive ability assessment to answer the following questions: (1) When given the opportunity to take practice tests, how often do candidates take them? (2) Who are the candidates that decide to take a practice test? What demographic differences are there in the uptake of practice tests by gender, age, ethnicity, neurodivergence & disability status?) (3) What impact does taking practice tests have on test score? (4) Are practice tests more effective for certain groups than others? This research hopes to inform whether more needs to be done to promote the uptake of practice tests, particularly amongst candidates from protected groups, or indeed, whether practice tests should be made mandatory to bring about improvements in adverse impact outcomes.

WEDNESDAY 3 JULY
Session 1.2.
Topic: International Assessment

102. Fostering Social Justice in South Africa through Collaborative Approaches to Enhance Cross-Cultural Assessments and Research

Naziema Jappie
University of Cape Town

This study explores the imperative of collaborative efforts to enhance cross-cultural assessment and research in South Africa, with a focus on promoting social justice. Rooted in a theoretical framework that draws from critical cultural psychology and intersectionality, addressing the complexities of cultural diversity. The objectives include identifying gaps in existing assessment practices, elucidating the impact of cultural nuances on research outcomes, and proposing collaborative strategies for improvement. The theoretical/conceptual framework integrates critical cultural psychology, emphasizing the need to go beyond superficial cultural considerations and engage with the underlying power structures and historical contexts that shape individuals' experiences. Intersectionality is employed as a lens to examine the interconnected nature of cultural identities in a diverse population. To achieve these objectives, a mixed-methods approach is employed that facilitate the exploration of participants' perspectives on existing practices and their implications for social justice. Preliminary results indicate a pressing need for improved cross-cultural assessment practices, as participants highlight instances of cultural insensitivity and biases affecting research outcomes. The implications of this study extend to the realms of academia, policy, and practice. By advocating for collaborative approaches, the research contributes to the advancement of social justice in South Africa. Recommendations include the integration of culturally sensitive research practices in academic curricula, the development of guidelines for ethical cross-cultural research, the establishment of collaborative networks to facilitate knowledge exchange and capacity-building. The findings call for a paradigm shift towards collaborative, culturally informed research practices that that promote inclusivity, fairness, and respect for the diverse voices within South Africa and the African continent.

WEDNESDAY 3 JULY
Session 1.2.
Topic: International Assessment

202. An Examination of School-Based and Work-Based Assessment Practices at China's Tertiary TVET Institutes

Gavin Brown, Yan Zhang, Jason Stephens
University of Auckland

In technical and vocational training, assessment practices in school-based contexts may differ considerably to those in work-based contexts. The former context conventionally emphasizes cognitive knowledge, while the latter prioritizes procedural competencies. Furthermore, formal school-based assessments may carry much more weight on final grades than the professionally judged competencies demonstrated in work-place internships. To examine these tensions, this study examined the school-based and work-based assessment practices from four selected disciplines (i.e., food science, health sciences, computer sciences, and mechanical/electrical engineering) at China's tertiary Technical and Vocational Education and Training (TVET) institutes. Structural Topic Modelling was applied to 166 assessment policies. Consistent with the traditions of formal summative assessment in China, but in contradiction to national policies, assessment practices tended to focus on summative, formal, percentage-scored examinations. Formative assessment and progress-related practices were absent, as were practical and internship judgment processes. Information on how portfolios, self-assessments, and third-party evaluations were to be conducted were largely absent. Small differences were detected through ANOVA between subject disciplines in terms of assessment practices. Consistent with the subject, food sciences had less emphasis on final exams than the others (d = .42) and had greater use of practical assessments over health sciences (d = .66). Health sciences emphasized test questions compared to the other disciplines (d = .44). These analyses show substantial weakness in institutional policies concerning the evaluation of practical, procedural, competencies needed for success in work. Clearly, breaking the power of the formal examination, notwithstanding China government policy, has still not trickled down into vocational and technical education.

WEDNESDAY 3 JULY
Session 1.2.
Topic: International Assessment

332. Item Quality for Cognitive Assessments in Low-to-Middle Income Countries: Evidence from the Ethiopia Young Lives Data

Winifred Wilberforce
Ohio State University
Anna A. O´Connell
Ohio State University, College of Education and Human Ecology

Conceptual framework This paper aims to examine how gender and location (rural vs urban) may impact item quality for assessments administered in low-to-middle-income countries (LMICs). We examine the cultural and gender-relevant items included in the cognitive assessments for the Young Lives school effectiveness survey data from Ethiopia (2016/2017). Objectives · Examines the item fit statistics and item targeting of the YL cognitive assessments for Ethiopia under the Rasch model. · Examines DIF analysis by gender and location for the cognitive assessment data. Sample The survey was completed by a total of N = 12,182 children in 7th and 8th grade from 63 schools. About half the sample identified as boys and 25.6% of the children were from rural Ethiopia. Methodology Item responses to the cognitive assessments were analyzed using the Rasch model. We conducted a DIF analysis to explore the possibility of item-level performance differences based on gender and location. Results The person distributions for both the Math & English assessments show that more difficult items should be added to the assessment for highly proficient students at the upper part of the scale. DIF Analysis For both English and Mathematics items, we used the Rasch-Welsh Test and a significance level of 0.05 based on the absolute value of the DIF contrast. There was no support for any individual items that favored boys or girls. However, for location, 15% of the mathematics items showed moderate to high DIF. The English test had 17.5% of items showing moderate to high DIF. For example, the first test item which asked students to identify a school bag, had a DIF contrast of .82. This item was harder for students in rural schools. Implications The literature from LMICs suggests that most large-scale test developers seem to focus on creating unbiased questions to avoid gender DIF than they do for location. We observe a similar trend in this study and hope to bring attention to this issue.

WEDNESDAY 3 JULY
Session 1.2.
Topic: International Assessment

574. Towards multi-cultural appropriate assessments in South Africa: Challenges with the test classification process

Justin August
Health Professions Council of SA
Deli Gumbi
Health Professions Council of South Africa

Dr Justin O August, Ms Deli Gumbi Health Professions Council of South Africa South Africa TOWARDS MULTI-CULTURAL APROPRIATE ASSESSMENTS IN SOUTH AFRICA: CHALLENGES WITH THE TEST CLASSIFICATION PROCESS South Africa's complex and controversial political history has not only impacted social and political spaces, but the ramifications of apartheid have influenced various academic disciplines. Psychological testing and assessment may be considered as one of the most affected areas within psychology. The majority of psychological tests were developed in Global north contexts and found their way to South Africa without adequately considering multi-cultural factors into their interpretations. The Health Professions Council of South Africa (HPCSA) Regulations defining the scope of Profession for Psychology reserves certain acts for the profession of Psychology, including psychological test use and publication of a classified list of tests. The process of classification prior to 2019 included an evaluation of cultural appropriateness amongst other areas. The objective of the paper is to highlight the test classification process in South Africa and the associated challenges. The regulatory position of the Professional Board for Psychology under the Health Profession Council of South Africa (HPCSA) will be explained. The presentation will further engage critical dialogue across regions with a potential for further engagement and collaboration on such processes. Keywords: cross-cultural appropriateness, Health Profession Council of South Africa, Psychological testing, psychological acts, regulations, test classification

WEDNESDAY 3 JULY
Session 1.2.
Topic: International Assessment

593. Pēhea Tōku Haerenga? A Māori Self-Assessment Measure for Aging and End of Life

Melissa Carey
University of Southern Queensland
Kathleen Mason
University of Auckland
Margaret Sandham
Massy University

Health and wellbeing measures utilized within Aotearoa New Zealand have been largely based upon ideas from the Western world. This has meant that for Māori people, the First Nations people of New Zealand, there is a mismatch between health and well-being needs and the tools used to assess these needs. A Kaupapa Māori community co-design approach was employed to gain understanding of the needs of older Māori people. The aim of the research was to develop a community framework, and toolkit, to support ageing and end of life care. A series of face to face and online workshops were conducted, participants highlighted the need for a tool that supported them to see how they "measure up" as they age. We analyzed Māori health models and existing measures to develop a Kaupapa Māori measure that enables older Māori people to make decisions about their ageing and end of life journey. We used a cultural design and a collective interactive approach to translate the measurement tool into an activity which can be used both online and offline. The design of the measure enables it to be a means for connecting individuals and groups to the resources that specifically relate to their personal needs. Enabling older Māori people to understand their ageing trajectory, connecting to resources and services that they need, has the potential to improve quality of life for older Māori people, and their families. The new measure enables older Māori people to connect with health services in a way that is meaningful to them, their families and the community.

WEDNESDAY 3 JULY
Session 1.3. SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

76. Examinee Engagement and Affect in Low-Stakes Testing Contexts: Influences and Personalized Interventions

ITC's (2013) Guidelines on Test Use call for test users to "consider other qualities which may have artificially lowered or raised results when interpreting scores" (p. 21). Four studies in this symposium show the importance of considering these qualities and examine strategies to influence them to achieve more accurate test scores. In the first study, we examined if examinee effort was influenced by how much effort students believe should be put forth on low-stakes tests (personal normative beliefs), beliefs about the effort level of other students (empirical expectations), and beliefs about what other students believe about effort (normative expectations). Beliefs were reported by 1144 U.S. undergraduates during low-stakes testing and related to self-reported effort, response time, and test performance, as expected. In the second study, 800 U.S. eighth graders were assigned to one of three conditions prior to a low-stakes math test: control, instruction, and nudge. Personalized nudges, but not instruction, reduced not-fully-effortful responses. Also, students often adjusted their effort level based on self-monitoring knowledge and effort. In the third study, 2367 U.S. undergraduates were randomly assigned to 1 of 5 motivation priming conditions before a low-stakes testing session: no priming questions, 3 intended effort questions, 1 intended effort question, 3 self-identity questions, and 1 self-identity question. Priming increased self-reported expended effort for male but not female students, with three questions having a larger impact than one question. In the fourth study, we examined the impact of providing immediate feedback to 410 German undergraduates during a low-stakes math test. Feedback benefited students' affective state after correct responses, but elicited unfavorable affective reactions after incorrect responses. Findings emphasize the need to account for performance when evaluating the potential of feedback for affective-motivational purposes.

WEDNESDAY 3 JULY
Session 1.3. SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

77. A personalized intervention study on decreasing not-fully-effortful responses during low-stakes mathematics assessment (Validity theory in testing, psychological assessment and survey research)

Burcu Arslan
Educational Testing Service Global
Bridgid Finn
Fidelity

In educational settings, students rely on metacognitive processes to determine whether or not to exert effort. It is important to detect and decrease not-fully-effortful responses during assessment because the inferences about students' knowledge, skills, and abilities are made with the assumption that they gave their best effort when taking the test. To this end, we investigated ways to minimize not-fully-effortful responses during a low-stakes mathematics assessment. Initially, we established theory-driven time thresholds for each item to detect such responses. We then administered the test to 800 eighth graders across three conditions: (a) control (n = 271); (b) effort instruction (n = 267); and (c) nudge (n = 262). In the effort instruction condition, students were told to exert their best effort before starting the assessment. In the nudge condition, students were prompted to give their best effort following each their first-attempt response that was both incorrect and not-fully-effortful. Thus, students had multiple opportunities to adjust their level of effort. Nudges, but not effort instruction, significantly reduced students' not-fully-effortful responses. Neither the nudges nor the effort instruction significantly impacted performance. In a posttest survey, most students reported that they received nudges whenever they did not know the answer (55%). Overall, these findings suggest that while nudges reduce not-fully-effortful responses, most students appear to strategically modulate their level of effort based on self-monitoring their knowledge and response effort.

WEDNESDAY 3 JULY
Session 1.3. SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

98. Perceived Normativity of Giving Effort on Low-Stakes Tests: Measures and Relations with Examinee Effort and Test Performance (Validity theory in testing, psychological assessment and survey research)

Dena Pastor
James Madison University/US
Sara Finney
James Madison University/US

In low-stakes testing, examinees have no personal consequences for test performance (Wise & DeMars, 2005). Thus, some examinees may expend minimal effort (O'Neil et al., 1995/1996). Factors likely contributing to effort levels, but yet to be investigated, include whether the examinee believes effort should be put forth (i.e., personal normative beliefs, PNB), how much effort they believe other examinees are putting forth (i.e., empirical expectations, EE), and their perceptions about what other examinees believe regarding giving effort (i.e., normative expectations, NE). In other words, examinee effort may be influenced by how much effort they think should be put forth on low-stakes tests, their beliefs about the actual effort level of other examinees, and their beliefs about what other examinees believe about effort allocation on low-stakes test. This logic is borrowed from economists and health researchers who have focused on using normative beliefs to change behavior in various domains: environmental conservation, corruption, charitable giving, sexual health, child marriage, alcohol consumption, and use of toilets (e.g., Bicchieri et al., 2023; Bicchieri et al., 2014; Constenbader et al., 2019). To inform norm-changing interventions, we need to understand individuals' beliefs about norms and how these beliefs impact behavior, which requires measures of these beliefs (Costenbader et al., 2019). In the current study, we created three measures of normative beliefs about giving effort on low-stakes tests, which were completed by a sample of 1144 US undergraduates during a large-scale, low-stakes testing session. All measures displayed desirable psychometric properties and were correlated but not redundant. Importantly, PNB and EE related positively to self-reported effort, response-time effort, and test performance, as expected. Although we advocate for assessing all three beliefs, PNB and EE may have the most utility for explaining effort during low-stakes testing.

WEDNESDAY 3 JULY
Session 1.3. SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

127. Exploring the Affective Impact of Immediate Feedback in Low-Stakes Testing: The Role of Student Performance and Visual Feedback Design (Validity theory in testing, psychological assessment and survey research)

Livia Kuklick
IPN Kiel, Germany
Marlit Lindner
IWM Tübingen, Germany

Immediate feedback in low-stakes testing has been proposed as a potential affective-motivational incentive to students, but its impact may not uniformly be positive. This experiment aimed to quantify the differential impact of positive and negative feedback in a low-stakes assessment with immediate performance feedback after each task. Additionally, the study aimed to explore ways to provide negative feedback in the affectively most beneficial manner. In a preregistered between-subjects study, 410 undergraduates participated in a computer-based geometry test. Participants were randomly assigned to one of five experimental conditions: (1) no feedback vs. immediate, elaborated feedback in one of four visual variants: (2) text only, (3) text + picture, (4) text + colors/animations, or (5) text + both additional design features. As expected, positive feedback after correct responses elicited higher levels of joy, hope, and pride, coupled with lower levels of anger and frustration. Conversely, negative feedback after incorrect responses resulted in inverted unfavorable affective reactions. The inclusion of one informative design feature (picture or colors/animations) mitigated the detrimental affective impact of feedback after incorrect responses compared to text-only feedback. However, these design features could not fully prevent test takers' detrimental affective reactions compared to receiving no feedback. This study highlights the importance of considering student performance as a determinant of affective reactions to feedback, illustrating the potential threat that negative feedback may pose to test score validity, given that test-taker emotions are important predictors of their performance. Moreover, it provides tentative evidence that a targeted visual design of feedback messages, incorporating informative features, may enhance the affective impact of immediate feedback in low-stakes contexts contributing to the ongoing discourse on effective feedback strategies.

WEDNESDAY 3 JULY
Session 1.3. SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

99. The Question-Behavior Effect in Low-Stakes Testing Contexts: How Many Questions are Needed to Prompt Good Test-Taking Effort? (Validity theory in testing, psychological assessment and survey research)

Sara Finney, Dena Pastor
James Madison University/USA

Questioning individuals about future behavior influences subsequent performance of that behavior (i.e., the question-behavior effect or QBE). Previous research found a QBE intervention significantly increased effort for undergraduates completing low-stakes tests (Finney & McFadden, 2023; McFadden, 2023). In the initial study (Finney & McFadden, 2023), students were randomly assigned to one of three QBE conditions prior to completing a low-stakes test: answering 5 questions regarding intended effort, answering 5 questions regarding intended effort that referenced positive self-identity, or no-question control condition. The subsequent study (McFadden, 2023) further supported the effectiveness of the intervention, but found 3 questions sufficient. In the present study, we examined if 1 priming question would be as effective as 3, as suggested by others (Wood et al., 2016). Moreover, because some strategies to increase effort have different effectiveness across gender (Braun et al., 2011), we examined gender as a moderator. We randomly assigned 2367 US undergraduates to 5 conditions prior to engaging in a low-staking testing session: no priming questions, 3 intended effort questions, 1 intended effort question, 3 self-identity questions, and 1 self-identity question. At the end of the session, students completed a measure of self-reported effort. We found a significant condition-by-gender interaction. Effort was not significantly different across QBE conditions for females, whereas for males effort was significantly lower for the no question condition compared to 3 of the 4 QBE conditions. Males also reported higher effort in the 3-question intended effort condition relative to both 1-question conditions. Similarly, the 3-question self-identity condition resulted in higher effort than the 1-question intended effort condition. Thus, we recommend 3 questions in QBE interventions and further study into the differential effects across gender.

WEDNESDAY 3 JULY
Session 1.4.
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

86. Language Cultural Considerations and Outcomes in the Translation of WRAT5 to Hebrew

Paul Shrell-Fox
Efrata Academic College
Dennis Bernstein, Avi Frank, Ayala Ahrell-Fox
PsychTech

Israel is a multi-cultural multi-lingual society. Languages spoken at mother-lounge levels include Hebrew, Arabic, English, Frech, Russin, Amharic and more. Within each of these language-cultures are numerous sub-sets of linguistic education. This is true of the Arabic as well as Hebrew speaking populations. The current project will present findings from the norming process of the WRAT-5HEB test of academic abilities. We embarked on the translation into Hebrew and adaptation to the Hebrew speaking residents of Israel of the test in 2018. We focused primarily on Hebrew given limited resources. (Our team began a broader project of translating the WPPSI-IV into both Arabic and Hebrew.) In the process of translation, we kept in mind the major language/educational cultures among the Hebrew speaking residents: public religious schools, public non-religious schools and ultra-orthodox schools. To be sure there are differences within these broad generalizations We predicted that on the two scales of reading, Word Reading and Sentence Comprehension there would be a specific advantage gained in reading development from within the ultra-orthodox schools. Given our experience as researchers/clinicians we predicted that this advantage would become smaller by middle school years (Grades 7-8; Ages 12-14). We predicted that there would be no significant differences in the Spelling or Mathematics test. At the time of writing, with preliminary data analyzed, the hypothesis has borne out to be true for the Word Reading sub-test. We have identified similar trends in the Sentence Comprehension test, though the analysis is not fully complete. By July 2024 all of the data will be available for presentation. In addition, as the norming process took place during the COVID-19 pandemic, a significant portion of the data was collected via teleconferencing. We predict no significant difference according to method.

These data will also be available by the time of the final paper presentation.

WEDNESDAY 3 JULY
Session 1.4.
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

239. Watch out for that item: Considerations on the linguistic adaptation of the Comprehensive Aphasia Test into Malay

Giuditta Smith
University of East Anglia
Mohd Azmarul A Aziz
Universiti Sains Malaysia
Maria Garraffa
University of East Anglia

T. Framework: Linguistic adaptation is a crucial aspect of the adaptation of evaluation instruments, particularly when what is being evaluated is language itself. The notion that linguistic stimuli must be adapted rather than translated has been one of the principles guiding the adaptation of the Comprehensive Aphasia Test (CAT, Swinburn, Peter, & Howard, 2004), a widely used assessment tool for language impairment in aphasia successfully adapted in 14 languages. Objectives: The active/passive sentence dichotomy (the boy was kissed by the girl/the girl kissed the boy) is employed in the English version of the CAT as a measure of sentence comprehension, with passive sentences being considered more complex, but adaptation guidelines specify other sentences must be employed when passives are infrequent in the target language (Fyndanis et al. 2017). In the current study, we aim to show that linguistic properties of the individual languages must also be considered when deciding which structures to include as reliable measures of comprehension. Methods: 14 healthy adults and 20 aphasics were tested on comprehension of reversible active and passive sentences on a sentence-picture matching task in standard Malay, a free word order language that relies on affixes for both active and passive sentences. Results: Aphasics were less accurate in the comprehension of reversible active and passive sentences, with no differences between sentence type. The active/passive dichotomy is therefore not a reliable measure of increased complexity in the comprehension of Malay. Implications: This study is part of an ongoing adaptation of the CAT in Malay, which is the first adaptation in an Austronesian language with free word order. Results confirm that, when adapting materials measuring sentence comprehension abilities in aphasic speakers of the language, it is crucial to be mindful not only of psycholinguistic measures such as frequency, but also of linguistic properties of the language.

WEDNESDAY 3 JULY
Session 1.4.
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

397. Psychological assessment in Brazil: Perspectives and Challenges

Solange Muglia Wechsler
Pontificial Catholic University of Campinas-Brazil

The history of psychological assessment in Brazil indicates there were four main phases of test development. During the first phase, the use of tests was mainly influenced by the movement in Europe and the United States; in the second phase tests were devaluated due to national criticisms as they were not representative of cultural characteristics; in the third phase tests were valued again due to the foundation of the Brazilian Institute of Psychological Assessment and the establishment of quality psychometric criteria for all tests to be approved. In the fourth current phase, new tests were constructed and validated to the country whereas foreign tests were rigorously adapted, validated, and normed to the Brazilian population. The criteria for test approval were based on ITC guidelines. The proportion of tests constructed and adapted from other countries will be presented. Although there was a significant development in the quality of test use in Brazil there are still many challenges to overcome. There is a relatively small number of tests for each area as they depend on the test editors 'investment. New technologies also call for different types of tests other than the traditional paper and pencil instruments on which Brazilians have a long way to prepare for test construction and adaptation. However, Brazil can be cited as a model for test development in South America.

WEDNESDAY 3 JULY
Session 1.4.
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

399. Challenges of psychological assessment: The case of Chile

Marcela Rodríguez-Cancino, Eugenia Vinet
Universidad de La Frontera

The current situation of psychological assessment in Chile is closely linked to social and historical events such as the proliferation of undergraduate training programs in psychology, the deregulation of professional practice, and the stagnation in the creation or adaptation of psychological tests in recent decades. This presentation exhibits a characterization of test users, identifying the most commonly used instruments and describing psychologists' attitudes toward their use. In addition, urgent challenges for the growth of this area of psychology are discussed, such as the need to strengthen the training of test users, the urgency of installing formal regulation mechanisms for the use of tests in the country, and the creation of lasting research teams dedicated to producing objective, expert, updated, periodic and easily accessible information, which will allow psychologists to make an informed selection of the tests to be used.

WEDNESDAY 3 JULY
Session 1.5.
Topic: Psychometric modeling

137. Latent growth analysis of serial eye-fixation indicators for multiple-choice test items

E. Cihat Corbaci
Sinop University
Nilufer Kahraman2/2
Gazi University

Built upon the Eye-Mind Theory relating eye movements to various cognitive processes, eye-tracking technology has been widely used in the field of educational testing. Unfortunately, item-related factors, such as task complexity, item difficulty and speed are too often shown to undermine the validity of the findings. The technology permits longitudinal model applications. However, this line of research has not yet been fully explored. To this end, this study considers several alternative latent growth model formulations to analyze eye-fixation measurements for the ultimate purpose of describing the overall within-item response processing trajectories and testing if examine ability (correct response) explains some of the variance. Application data were item response and eye-fixation measurements (250 Hz, Smart Eye) of 159 college students during a ten-item computerized multiple-choice test experiment (measuring reading comprehension skills in English). First, to handle the serial dependency among the within-item segments (directions, text lines, and choices), the measurements were binary coded into a series of longitudinal indicators (AOIs). Next, the cumulative fixation durations (in Sc), computed using each item's respective AOI series, were modeled. Results show that a general curvilinear pattern approximates the overall within-item processing patterns the best, i.e., the average processing times increasing steadily at first (from the first AOI to the next), and then dropping slightly at the end. The results reveal that both examinee ability and item characteristic (difficulty, the placement of the correct-coded choice, poor distractors) might influence the within-item trajectories. These findings suggest that longitudinal models might be useful for further exploring, for example, if examinees change or adapt their response processing behavior depending on item types, or if some item types or response processing patterns might be more conducive to speed effects. This study was partially supported by TUBITAK under grant SOBAG 120K142.

WEDNESDAY 3 JULY
Session 1.5.
Topic: Psychometric modeling

148. Psychometric evaluation of the environmental knowledge scale in TIMSS 2019

Purya Baghaei, Yuan-Ling Liaw, Rolf Strietholt, Sabine Meinck, Andrés Strello
IEA-Hamburg

This study aims to review and investigate the environmental knowledge scale of the TIMSS 2019 science assessment. We put particular focus on the distinctness of general and subscale scores, including the newly introduced environmental knowledge scale. In this study, we review the TIMSS science assessment framework and its psychometric scaling procedure. We also outline the process of constructing the TIMSS 2019 environmental knowledge scale, including the definition of the construct, the identification of environmental items, and the scaling of environmental awareness data. Finally, the dimensionality of the TIMSS 2019 science assessment is evaluated. It was specifically tested whether the data can be explained by a unidimensional IRT model or a bifactor model. Examining information criteria, subscale reliabilities, and explained common variances we found evidence supporting both the existence of a strong general science factor and nested factors for environmental knowledge and other science subdomains. Implications of the findings for the TIMSS subscale reporting are discussed.

WEDNESDAY 3 JULY
Session 1.5.
Topic: Psychometric modeling

257. An Item Response Tree Model for Items with Multiple-Choice and Constructed-Response Parts

Junhuan Wei, Yan Cai, Dongbo Tu, Zhichen Guo, Qin Wang, Kai Liu, Daxun Wang, Fen
Luo, Fangbin Chen, Jiyuan Ding, Xuhong Song, Pan Jiang
China

Multiple-choice (MC) items and constructed-response (CR) items are common forms in large-scale tests. Recently, a new item format has emerged in educational and psychological measurement, incorporating a blend of MC and CR items. In this innovative format, candidates are prompted to initially choose an option within the MC task, followed by responding to questions associated with the selected option in the CR task. Traditional IRT and IRTree models are not appropriate for analyzing the item that simultaneously consists of MC task and CR task in one item. To address this issue, this study proposed an item response tree model (called as IRTree-MR) to accommodate items that contain different response types at different steps and multiple different cognitive processes behind each score to effectively investigate the cognitive process and achieve a more accurate evaluation of examinees. The proposed model employs appropriate processing function for each task and allows multiple paths to an observed outcome. The simulation studies were conducted to evaluate the performance of the proposed IRTree-MR, and results show the proposed model outperforms the traditional IRT model in terms of parameters recovery and model-fit. Moreover, an empirical study was carried out to verify the advantages of the proposed model.

WEDNESDAY 3 JULY
Session 1.5.
Topic: Psychometric modeling

424. A Latent Profile Analysis of Examinees' Multiple-Choice Item Processing Behavior Using Segment-Specific Eye-Fixation Metrics

Derya Akba
Aydın Adnan Menderes University, Turkiye
Nilufer Kahraman
Gazi University, Turkiye
Ergun Cihat Corbaci
Sinop University, Turkiye

Eye-tracking technology is increasingly used in educational assessment to study qualitative individual differences in item processing behavior of examinees. Previous studies have presented promising findings, but more research is needed to examine the item-related factors that might undermine the consistency and validity of the results. This study uses latent profile analysis to evaluate the overall within-item over-segment (Area of Interest) eye movement patterns of a group of examinees using their item responses (0/1) as a covariate variable. Application data were from 152 college students responding to a five-item reading comprehension test in English and included item responses (0/1) and eye-fixation metrics (Smart Eye, 250 Hz). Firstly, segment-specific (average log) process times (in sec) were computed for each item using a series of indicator variables, marking item segments, starting from the directions at the top (1), followed by the lines in the stem (2 to 5), and the choices at the end (6 to 10). Then, a series of latent profile models (from one-class to four-class models) were tested for each item. The results revealed that the three-class model had the best fit across all the items. These classes were labeled as low, medium, and high-effort groups considering their mean item process times. Moreover, the findings indicated that the examinees with incorrect responses were more likely to be in the moderate and high-effort groups rather than the low-effort groups (shorter item encounter times). The results of the study suggest that latent profile models can be a valuable tool for assessing qualitative individual differences in item processing behavior and to study the influences of various item features (e.g., item type, difficulty) on item processing patterns. This study was partially supported by Gazi School of Education and by TUBITAK under grant SOBAG 120K142.

WEDNESDAY 3 JULY
Session 1.6. SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

153. Changes in Collegiate Admissions Policies and Procedures

Kurt Geisinger
Buros Center for Testing, University of Nebraska-Lincoln, USA
Wayne Camara
Law School Admissions Council, USA
Maria Elena Oliveri
Buros Center for Testing, University of Nebraska-Lincoln
Discussant name
Kadriye Ercikan
Discussant surname
Ercikan
Discussant affiliation
Educational Testing Service

This symposium will provide three papers from distinguished members of the admissions testing community. First, Professor Maria Elena Oliveri from the Buros Center for Testing at the University of Nebraska-Lincoln will present "Global Perspectives on Higher Education Admissions: Navigating Access, Diversity, and Equity Challenges." In her presentation, she will describe differences in admissions policies around the world as found in her ITC published volume, Higher Education Admissions Practices: An International Perspective on admissions testing and processes around the world. Next Wayne Camara 's (from the Law School Admissions Council) paper will describe policy changes that U.S. colleges and universities have made and show how such policies impact the diversity of admitted and enrolled students and describe initial findings from institutions. Then Kurt Geisinger, Director of the Buros Center for Testing at the University of Nebraska-Lincoln will describe the broader context of why colleges and universities made these procedural changes, including the advancement of diversity, and how many colleges and universities are now making admissions decisions. Both Camara and Geisinger will describe some recent validation efforts demonstrating the effects of the traditional admissions tests. Finally, Kadriye Ercikan has agreed to serve as a discussant for this session.

WEDNESDAY 3 JULY
Session 1.6. SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

155. Changes in Admission Tessting Practices in Colleges and Universities (Translation of tests, psychological assessment instruments and survey questionnaire)

Kurt Geisinger
United States

The volume and use of college admissions testing has changed dramatically in the past five years. Many (about 80%) institutions now do not require test scores for admissions. These changes have occurred due to the pandemic, the Black Lives Matter movement, and the Varsity Blues scandal. This paper evaluates the benefits and costs of these changes and whether they are likely to be longstanding. Some recommendations are also made for ways to increase the diversity of student bodies while maintaining maximal validity. In addition, it looks at ways that non-test information such as grades can be used in a manner that is most equitable. The paper will also present results of the recent University of California validity studies of the SAT and ACT and grades with implications for how admissions offices should use such information in the future most validly and fairly manner.

WEDNESDAY 3 JULY
Session 1.6. SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

163. Global Perspectives on Higher Education Admissions: Navigating Access, Diversity, and Equity Challenges (International assessment)

Maria Elena Oliveri
Buros Center for Testing

The presentation, based on the ITC book "Higher Education Admissions Practices: An International Perspective," edited by Oliveri and Wendler, delves into the intricate challenges confronting global higher education admissions, focusing on diversity, equity, and access. Drawing upon theoretical and conceptual models from various countries worldwide, this Session provides concrete examples from diverse institutions of higher learning facing challenges such as access to quality education, promotion of diversity, and pursuit of equity reflect a broader commitment to ensuring educational opportunities for a wide demographic. The presentation not only outlines existing challenges but also illuminates the current global landscape of higher education admissions practices. By examining examples and theoretical model from various countries (e.g., the multilevel design and complementarity design for assessment development and use), the presentation goal is to provide a nuanced understanding of the diverse approaches and strategies employed by researchers and educational institutions to address these challenges. Moreover, the presentation explores collaborative efforts among researchers across different countries, delving into shared insights and solutions devised to address complexities in higher education admissions. Through a comparative lens, the audience gains valuable insights into successful strategies in diverse cultural and institutional contexts. Essentially, this presentation serves as a platform for the exchange of knowledge and best practices, fostering a global dialogue on higher education admissions. By showcasing examples from different nations, it provides a comprehensive understanding of the dynamic and evolving nature of admissions practices, contributing to the collective effort to enhance access, diversity, and equity in higher education on an international scale.

WEDNESDAY 3 JULY
Session 1.6. SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

174. Test optional policies: trends and impact - to date (Validity and fairness in cross-cultural testing, psychological assessment and survey research)

Wayne Camara
LSAC

This presentation will focus on the impact of test optional and test blind practices and trends in US admissions testing. Admissions and admissions testing has undergone profound changes since 2019. Prior to the Pandemic, 12 percent of four-year colleges didn't require tests, representing 1% of all freshmen (Camara, 2020). For 2023, 80 percent of four-year colleges do not require the ACT or SAT for admission and over 80 colleges being test-blind (FairTest, 2023). A number of state colleges and independent institutions extended test-optional policies through 2023, 2024 or indefinitely and test-optional policies are increasingly common within graduate admissions. This year, 48% of applicants submitted a test score with their college applications, up from 44% last year, but down from 76% in 2019–2020 (Freeman, Magourik, and Kajikawa, 2021). A Task Force established at the University of California strongly rejected a test-optional policy, which was subsequently approved by the UC President and Regents. The Task Force found "with some confidence, we can predict that the following would happen if UC stopped using admissions tests and relied solely on (high school) GPA and other aspects of the students' transcript for the academic appraisal of applicants" (2020, p. 85): • The average student admitted to UC would have a lower FGPA, probability of persistence, probability of graduation, and lower GPA upon graduation. • The average financial subsidy by the state would have to rise due to the longer time to graduate per student and costs per bachelor's graduate would rise due to extended time to graduation and the increased dropout rate. • UC would have less ability to identify and target high-risk students for academic support without tests. Additional research is needed, but only with serious studies that control for differences in test submitters and non-submitters, or applicants/ matriculants in 2021-23 vs those in prior years.

WEDNESDAY 3 JULY
Session 1.7.
Topic: Testing equivalence by psychometrics methods

108. Measurement invariance/equivalence of the Index of Psychological Well-being at Work Across Black and White South African employees

Gina Görgens-Ekermans, Lara Kelly
Stellenbosch University

Set within the unique legislative context in South Africa that governs psychological testing, this study addressed the cross-cultural suitability of the Index of Psychological Well-Being at Work (IPWBW) (Dagenais-Desmarais & Savoie, 2012) over Black and White South African employees, through a measurement invariance/equivalence (MI/E) investigation. The legal framework (Employment Equity Act 55 of 1998) governing testing in South Africa requires psychologists to proactively provide evidence that psychological tests are valid, reliable, fair and unbiased. Cultural, as well as measurement artifacts, may result in a lack of measurement invariance across different groups (e.g., race, language). Therefore, this study aimed to assess the MI/E of the IPWBW over Black and White respondents, by applying the Dunbar et al., (2011) MI/E taxonomy. A cross-sectional archival dataset (N = 366) with 51.4% Black and 48.6% White employees were utilized. The MI/E procedure requires a series of multi-group CFA models to be fitted to the data. Configural invariance was not found and further MI/E analyses were terminated. Consequently, EFA, using RStudio, with orthogonal procrustean target rotation was conducted to calculate a factor congruence coefficient, Tucker's Phi. The results revealed good (White; $\alpha \geq 0.70$) to reasonably good (Black; $\alpha \geq 0.67$) internal reliability for all five subscales of the IPWBW. The individual group CFA results for the Black sample revealed that for 96% of the items, significantly more variance was explained by measurement error than the dimensions the indicators were tasked to represent. In addition, lack of configural invariance was found, suggesting the presence of construct bias, further corroborated by the coefficient congruence matrix (

WEDNESDAY 3 JULY
Session 1.7.
Topic: Testing equivalence by psychometrics methods

113. Measuring Personality in a Changing World: Psychometric Analyses of the BFI-2 across Offline and Online Situations

Dora Leander Tinhof, Axel Mayer
Bielefeld University

Ensuring reliable and valid measurements of psychological constructs is not only crucial for enhancing replicability and generalizability in research, but also for ensuring meaningful interpretability of results. Notably, even instruments designed to assess stable constructs like the Big Five personality traits exhibit substantial variability across situations (Baumgartner & Steenkamp, 2006; Geiser et al., 2015). Recognizing the increasing relevance of the digital world alongside the "real" world (Kaufmann et al., 2020; Schwarz et al., 2020), it becomes crucial to explore the functionality of Big Five trait measurement instruments in both online and offline situations. To this end, this study employed a longitudinal multi-rater and multi-situation design to collect self- and other-report data, utilizing the German Big Five Inventory-2 (Danner et al., 2019), across offline and online situations at two measurement timepoints. Initially, the psychometric properties, dimensional structure, and measurement invariance were examined across all combinations of methods and situations. While generally demonstrating good internal consistencies at facet and domain levels, item reliabilities varied more substantially within domains than across situations. The original dimensional structures did not consistently align with the data. While self-report measures generally achieved at least metric invariance and performed better than other-report measures, no systematic offline – online differences emerged. Using multi-method latent state-trait models for random and fixed situations (Hintz et al., 2018), we subsequently disentangle stable and variable measurement components and identify person × situation interaction effects under consideration of method effects (target N = 500; data collection ongoing). The discussion highlights the most notable findings and addresses their implications for both research and practice.

WEDNESDAY 3 JULY
Session 1.7.
Topic: Testing equivalence by psychometrics methods

390. Esra Sözer Boz (Bartin University/Turkey) Alignment Optimization to Test Measurement Invariance of Mathematics Anxiety across 36 Countries

Aishwarya Jaiswal, Akansha Tyagi, Rob Bailey, Himanshi Sharma, Karishma Agarwal Mercer | Mettl

553 Karina da Silva Oliveira (Universidade São Francisco), Ana Clara Silva Resende (Universidade Federal de Minas Gerais), Clara Paes Silva (Universidade Federal de Minas Gerais), Franciele Neves Moreira (Universidade Federal de Minas Gerais), Ana Carolina Cordeiro Alves(Universidade Federal de Minas Gerais), Carolina Guitzel Borghi (Universidade Federal de Minas Gerais), Geovana Silva (Universidade Federal de Minas Gerais), Jade Tavares Pereira Liberato (Universidade Federal de Minas Gerais), Thalita Cezar Aguiar (Universidade Federal de Minas Gerais) Children's Resilience Markers: Initial Studies for Age Range Expansion

Children's Resilience Markers (CRM) assesses resilient behavior in Brazilian children aged eight to 12 years. This 22-item test, uses situational judgment format, and has undergone investigation across various conditions, accumulating evidence suitable for the proposed age range. Due to the scarcity of instruments for child assessment in Brazil, especially for those under eight years old, this study aims to investigate the possible replicability of the CRM's factor structure in children aged five to 12 years. Participants included 1005 Brazilian children (girls=443, boys=562), aged five to 12 years old (M=9.53, DP=1.79), from three Brazilian States. The original 22-item CRM format was utilized. Each item presents a brief story where the main character faces a typical child's daily life situation. The story halts, prompting the character to decide how to resolve the challenging situation. And the child must choose the option that best represents the child's behavior. Data analysis used FACTOR statistical package, employing Exploratory Factor Analysis with factor extraction through Parallel Analysis, Oblique rotation, and Principal Axis Factoring estimation method. Initial data factorability assessment yielded KMO=0.89; Bartlett p

WEDNESDAY 3 JULY
Session 1.8.
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

47. Integrating process data with response data for cross-cultural comparability and data insights

Hongwen Guo
ETS

With technology advances, digital-based large-scale assessments (LSAs) can capture test-taking process data alongside task responses on their platforms, offering insights into students' cognitive thinking processes, time management, test strategies, etc., beyond a performance score. This wealth of information holds great promise for researchers, educators, and policy makers, as it enables a deeper understanding of how students from different cultural-language groups perform on and engage with LASs differently. The objective of the presentation is to discuss several statistical methodologies to analyze data from multiple sources for group comparison, demonstrate their applications on large samples of students collected from LSA (such as PISA), and their relevance to test validity and comparability across different language-culture groups. Statistical methodologies for group comparison, such as differential item functioning (DIF), and its extension to process features (such as response time, students' interactions with the platform) are discussed, as well as statistical methods for detecting rapid guessing behaviors that may compromise score validity and comparability. Results show that test-taking behaviors are not a nuisance factor that may confound measurement but an aspect providing important information on how students approach tasks. They may have different relationships with performance for different groups with different pre-knowledge, backgrounds, languages, social-cultural norms, and learning and assessment experiences. Ignoring test-taking differences might lead to problematic group comparison and difficulty in data interpretation. Findings based on these methodologies highlight the importance of taking cross-cultural differences into account when interpreting the rich process data from international LSAs and when providing feedback for different stakeholders.

WEDNESDAY 3 JULY
Session 1.8.
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

408. Threats to validity and fairness of testing: how different demographic and cross-cultural groups experience testing differently

Stephen Cuppello
Thomas International

In the context of employee selection, psychometric assessments not only provide excellent predictive validity for future performance but also play a pivotal role in enhancing the fairness of decision-making processes by providing a source of objective information. Despite this, there is evidence that different groups of people experience tests differently, and that these differences may introduce measurement bias. Following a systematic review of existent literature, three factors with empirical support were identified as potential sources of bias: stereotype threat, test confidence, and variations in test and testing domain experience. The primary objective of this study is to investigate national, gender, age and socioeconomic status (SES) differences in experiences relating to these factors. Some 934 participants were recruited from two distinct cultures: the UK and South Africa. To focus on perceptions of real occupational test takers, participants completed a perception, experience and demographic survey following completion of a psychometric assessment for genuine occupational test use. Statistically significant differences in test taker experiences were found across all groups. In identifying these differences, this research not only contributes to the ongoing discourse on validity and fairness in testing cross-culturally but also provides insights into potential areas for improvement in the design and administration of psychometric tests. As we strive for equitable and inclusive decision-making processes, understanding the nuances of these factors becomes paramount. This research has broader implications for test developers, administrators, and policymakers seeking to enhance the validity of assessments on a national and international scale. By identifying and addressing the specific elements that introduce biases, we take a step forward in ensuring that decisions based on test scores contribute positively to equity, inclusion, and fairness.

WEDNESDAY 3 JULY
Session 1.8.
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

557 Aishwarya Jaiswal (Mercer | Mettl), Akansha Tyagi (Mercer | Mettl), Rob Bailey (Mercer | Mettl), Himanshi Sharma (Mercer | Mettl), Karishma Agarwal (Mercer | Mettl) Fortifying the Human Firewall: Development and Validation of a Personality-Based Organizational Cybersecurity Risk Assessment Framework

WEDNESDAY 3 JULY
Session 1.8.
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

748. Institutional Diversity and its Implications for Assessments and Outcomes

Daniel McCaffrey, Heather Buzick, Jonas Bertling
ETS
Indrani Bhadur
NCERT
Mariya Khatoon
ETS

Indian school boards vary in their assessment practices and governance approaches. This study provides a snapshot of the variability so that stakeholders make informed decisions in establishing equivalence guidelines for school boards in the nation. School boards play a pivotal role in the Indian educational system with oversized impact on post-secondary success and outcomes; the lack of equivalence could lead to differential outcomes for students. This work is part of a larger effort to improve opportunities for all students to learn and demonstrate competencies within a system of equivalent academic standards while integrating national and local culture. To this end, we developed, piloted and implemented a tool to record assessment and item details, including an item inventory, to capture information about the board examinations. Development was informed by best practices in assessment and item review. We paired the tool with publicly reported board examination participation and performance data to answer two main research questions: What are the similarities and differences in what and how boards currently assess? How do examination participation and passing rates vary within and between education boards? 51 school boards were invited to participate, representing all secondary, higher secondary, and open school boards in the nation. A diverse set of 36 boards submitted data, which included ratings of a sample of 138 tests (approximately 6,000 items) on several content areas; analysis focused on mathematics and English, which had the most ratings. The publicly available annual data on secondary and higher secondary exam participation and performance by boards spanned years 2009 to 2022. Using descriptive statistics, correlational analyses, and linear mixed-effects modelling, we demonstrated large variabilities across school boards in (a) assessment specifications (item types, difficulties, cognitive demands), (b) gender participation ratios, and (c) passing rates.

WEDNESDAY 3 JULY
Session 1.8.
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

787 Mixed Item Type Strategies to Support Culturally Relevant Cross-Cultural Measurement

Fernanda Gandara
Room to Read

One of the dilemmas in international assessments is related to the issue of score equivalence. Psychometricians have described the levels of equivalence that one can achieve and their implications for analysis. Typically, professionals refer to construct equivalence –where an instrument is measuring the same trait across contexts; measurement equivalence –where measures also have the same unit; and scalar equivalence – where measures across contexts also have the same origins (van de Vijver & Tanzer, 2004). Score equivalence needs to be established and not assumed. The problem of generating comparable scores in international assessments is heightened by the call for more culturally relevant measurement. Culturally relevant measurement refers to creating measures that are appropriate and important to the examinees and local stakeholders (Casillas, Roberts, & Jones, 2023). Culturally relevant measurement demands a higher emphasis on context, which jeopardizes the overlap of constructs and the ability to establish meaningful comparisons between contexts. The purpose of this study is to explore a novel strategy to deal with the tension between cultural relevance and comparability in international assessments. The strategy involves developing measures for the same constructs that include both global and local items, a.k.a. mixed item type strategy. The assumption is that the first type should yield comparable scores, focusing on shared portions of constructs, and the second type may or may not, depending on the context(s). The research questions that we aim to respond are: a) What are the necessary technical considerations for this strategy to support appropriate inferences within and across contexts? b) What are the benefits and limitations of this strategy compared to traditional adaptation procedures? The results of this study have implications for multi-country programs who deal with these dilemmas on a constant basis.

WEDNESDAY 3 JULY
Session 1.9.
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

100. Self-assessment scale of emotional labour for early childhood teachers: A context-centered approach

Qilong Zhang
United Arab Emirates University/United Arab Emirates
Jianqin Yin
Jiangsu Second Normal University/China

Emotional labour is an important part of teacher wellbeing and teaching effectiveness. Teachers work in a context that is different to that of other service sectors, which may lead to different emotional labour strategies, and early childhood teachers in particular. Adopting a context-centred approach and utilising the context of mainland Chinese kindergarten teaching, this study presented an emotional labour self-assessment tool for professional learning among early childhood teachers. With three convenience samples (three phases) of a total of 1020 kindergarten teachers in China, this study identified typical emotional labour situations centred on which the self-assessment tool was developed and validated. The tool included 25 items, with six items of surface acting, nine items of deep acting, and 10 items of expression of naturally felt emotions (ENFE). Three patterns of early childhood teachers' emotional labour were identified: (1) There was an apparently higher mean frequency of surface acting compared to deep acting and ENFE; (2) There was no negative relationship between surface acting and deep acting or ENFE, indicating the absence of a good versus bad dichotomy of emotional labour; (3) Deep acting and ENFE were highly correlated, indicating the absence of a clear boundary between deep acting and ENFE. Different to previous studies in which decontextualized scales of emotional labor were used, this study highlighted the importance of the context in understanding and measuring emotional labor among teachers.

WEDNESDAY 3 JULY
Session 1.9.
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

136. Cognitive and Emotional Processing of Culturally Responsive Test Items

Chris Patterson
University of Iowa

The call for tests to be more culturally responsive or antiracist includes inputting topics like genocide and pregnancy into test items; topics that traditional test development lenses see as controversial. As a result, there is a concern from those in traditional test development that topics discussing culture or racism will induce strong cognitive confusion or emotional reactions, which can invalidate test scores. Although these concerns are rooted in racism and white supremacy, these concerns point out a need to understand the cognitive and emotional processes of those answering items with racialized contexts. The objective of this study was to more directly addresses concerns through determining if and how test takers cognitively processed and emotionally reacted to culturally responsive and antiracist item topics. A maximally varied sample of 20 college students were recruited to participate in 1-hour cognitive interviews, each interview consisting of 10 items containing culturally responsive or antiracist topics. Thematic analysis of participants' response processes were heavily characterized by the interaction of race of the participant and type of topic interacted with. While students of color approved of and felt joy and happiness while processing both culturally responsive and antiracist items, white students felt cognitive dissonance while answering antiracist items. Further, some students of color experienced strong negative opinions toward the inclusion of specific topics in items. Implications primarily call for replication studies with more test taking populations and different topics discussed in items while reinforcing the need for a multicultural group of item developers to create items that can discuss appropriate topics on tests that won't deface test takers' of colors' cultures while introducing a healthy amount of racial dissonance to balance answering items in accordance with test takers' true ability and healthy racial identity development.

WEDNESDAY 3 JULY
Session 1.9.
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

648. Towards the Development of a Tool to Measure Socio-Emotional Competencies

Theresse Dela Cruz, Arnelli Atentar, Jillian Navarrete, Rose Arvie Dela Cruz, Shiena Go, Rose Ann Competente, Nicolle Anne Yabut
Global Resources for Assessment Curriculum and Evaluation, Inc.

Socio-emotional learning (SEL) refers to developing essential noncognitive skills beneficial to students. Although the importance of SEL has been recognized globally, only a few local studies have explored how these factors can contribute to our understanding of Filipinos as learners. In response, the researchers designed a tool to measure socio-emotional competencies through the following scales: self-awareness, self-management, self-regulation, social skills, and grit. The SEC tool has four forms with statements worded in English and translated into Filipino: Primary, Intermediate, Junior High School, and Senior High School. Responses from 1,211 Filipino students who participated in the pilot tests were analyzed. Based on the conducted equivalence study, exploratory factor analysis, and reliability analysis, the internal consistency of SEC forms is high (r=.76-.94). The resulting factors show good model fit (TLI=.91-.97 and RMSEA=.02-.04). Predictive analyses suggest that the level of socio-emotional competence is significantly different across students' grade levels; results show that Filipino students at higher grade levels are lower in social-emotional competence compared to students at lower grade levels. Researchers recommend increasing the number of items on the scales with weak internal consistency and conducting correlational studies examining the relationship of socio-emotional competencies with students' academic performance.

WEDNESDAY 3 JULY
Session 1.9.
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

753. Assessing social-emotional competence in teacher training admissions

Florian G. Hartmann, Tuulia Ortner, Andreas Kurz
Paris Lodron University Salzburg
Poster

This research deals with properties of a new test assessing aspects of social-emotional competence in an admission procedure for teacher training in Austria. Given the importance of social skills in the educational context, the social skills of prospective teaching students are crucial. We present the psychometric properties of a new test developed from expert interviews, consisting of school-related social situations. Within this new approach, test takers were not only instructed to indicate appropriate emotions demonstrated by teachers in scenarios, but were also, afterwards, instructed to indicate the likely emotions of the students in the situations. The analyzed data consisted of N = 1977 individuals who took the electronic test between 2021 and 2023 as part of an admission procedure for teacher training in a part of Austria (secondary general education). In addition to this social-emotional competence test, participants were asked to complete other assessments (e.g., a test on dealing with numbers), utilized in the present study to examine discriminant validity. Strengths of the test as, for example, the embedding in realistic social situations as well as possible weaknesses (e.g., ceiling effects, scoring) are discussed in the light of the aim of the selection procedure.

WEDNESDAY 3 JULY
Session 1
Topic: Artificial Intelligence in testing, psychological assessment and survey research

725. Enhancing ADHD Screening in Children: Integrating Machine Learning with Item Response Theory.

Alexandre Serpa
Mackenzie Presbyterian University, São Paulo, Brazil
Pedro Loures Alzamora, Derick Oliveira, Camila Nicola, Victoria Oliveira, Laura Ludgero, Ana Paula Couto Silva, Gisele Pappa
Universidade Federal de Minas Gerais, Brazil
Marco Romano-Silva
Centro de Tecnologia em Medicina Molecular (CTMM) da UFMG, Minas Gerais, Brazil
Wagner Meira Jr
Universidade Federal de Minas Gerais, Brazil
Débora Marques Miranda
Centro de Tecnologia em Medicina Molecular (CTMM) da UFMG, Minas Gerais, Brazil

Attention Deficit Hyperactivity Disorder (ADHD) affects approximately 5% of children and 2.5% of adults. In Brazil, less than 20% of Brazilians with ADHD receive adequate treatment, with marginalized communities particularly underserved. This study aims to develop an accurate and interpretable ADHD classifier for children and adolescents by integrating machine learning (ML) and item response theory (IRT). The study analyzed data from 345 children and adolescents assessed at the Universitary Hospital of Federal University of Minas Gerais using the K-SADS-PL diagnostic method and seven psychometric assessments. Collected between 2011 and 2020, the data include 23 ADHD-related variables covering patient and family information, parental styles, and socioeconomic context. The methodology involved four steps. Initially, data homogenization was done by standardizing null values and binarizing test items. Sensitive attributes were removed for ethical compliance. The dataset was then split into training and test sets. Four decision tree-based classification models - Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Explainable Boosting Machine (EBM) - were compared using Random Search optimization and k-fold cross-validation. The EBM model showed superior performance (F1-score = 0,87, recall = 0,98, specificity = 0,14, precision = 0,78, accuracy = 0,77) and selected 84 features including sociodemographic, parental styles and test items. Rasch model indicates good reliability and fit index for most items. Theta ranges from -4,010 to 2,858 (M = 0,280, SD = 1,515). Item's infit ranges between 0,846 and 1,186 and outfit ranges between 0,698 and 1,886. The separation index was 0,835, indicating a high reliability. Results demonstrate the potential of integrating ML and IRT to enhance ADHD screening in children. Findings suggest promising avenues for early diagnosis and intervention, potentially improving outcomes for children with ADHD.

WEDNESDAY 3 JULY
Session 1
Topic: Artificial Intelligence in testing, psychological assessment and survey research

432. AI for Psychometrics: Validating Machine Learning Models in Measuring Emotional Intelligence with Eye-Tracking Techniques.

Chapman Lindgren
The Graduate Center & Baruch College, CUNY
Wei Wang
The Graduate Center, CUNY
Liat Koffler
The Graduate Center & Brooklyn College, CUNY
Max Lobel, Amanda Murphy, Qiwen Tong, Kemar Pickering
The Graduate Center, CUNY

AI, or artificial intelligence, is a technology of creating algorithms and computer systems that mimic human cognitive abilities to perform tasks. Many industries are undergoing revolutions due to the advances and applications of AI technology. The current study explored a burgeoning field—Psychometric AI, which integrates AI methodologies and psychological measurement to not only improve measurement accuracy, efficiency, and effectiveness but also help reduce human bias and increase objectivity in measurement. Specifically, by leveraging unobtrusive eye-tracking sensing techniques and performing 1470 runs with seven different machine-learning classifiers, the current study systematically examined the efficacy of various (ML) models in measuring different facets and measures of the emotional intelligence (EI) construct. Our results revealed an average accuracy ranging from 50–90%, largely depending on the percentile to dichotomize the EI scores. More importantly, our study found that AI algorithms were powerful enough to achieve high accuracy with as little as 5 or 2 s of eye-tracking data. The research also explored the effects of EI facets/measures on ML measurement accuracy and identified many eye-tracking features most predictive of EI scores. Both theoretical and practical implications are discussed.

WEDNESDAY 3 JULY
Session 1
Topic: Artificial Intelligence in testing, psychological assessment and survey research

610. Exploring the Reliability of Audio Signals in Video Interviews for the Automatic Prediction of Psychological Characteristics.

Borja Artiñano, David Aguado, Pablo Garcia, Sara Estirado
Instituto de Ingeniería del Conocimiento

Asynchronous Video Interviews (AVIs) are widely utilized in the field of personnel selection to assess candidates for job positions quickly and cost-effectively. Alongside this, technological advancements in mathematical-computational developments associated with Machine Learning and Artificial Intelligence enable the automatic analysis of AVIs and the estimation of candidates' psychological characteristics predictive of their future professional performance (e.g., personality). However, despite their extensive use, there is limited evidence regarding the psychometric properties of automatic estimations on the content of AVIs. The objective of this study is to analyze the reliability of one of the signals used by Artificial Intelligence to make these automatic inferences: the audio signal. This signal refers to the speaker's voice characteristics and is commonly termed prosody. The study analyzed the audio signal from a set of AVIs (N=39). Audio features available in the eGeMAPS model were extracted using the OpenSmile library, and scores obtained for different audio features extracted across different time segments were compared. The results indicate that there are audio features extracted from AVIs that are more stable than others and, therefore, cannot be equally utilized in automatic prediction models. In audio segments between 30 and 40 seconds, a higher number of stable features over time seem to be present. These findings have significant implications for the design and development of automatic systems predicting psychological variables through the audio signal recorded in AVIs.

WEDNESDAY 3 JULY
Session 1
Topic: Artificial Intelligence in testing, psychological assessment and survey research

73. Distilling vector space models for psychoeducational assessment: honing semantic indicators in automated summary evaluation.

José Ángel Martínez-Huertas
National Distance Education University
Guillermo Jorge-Botana
Complutense University of Madrid
Ricardo Olmos, José A León
Autonomous University of Madrid

Computational semantic measures from vector space models are relevant to obtain indicators of different psychological constructs. In the evaluation of constructed responses, as in automated summary evaluation, text responses are represented in the vector space. The coordinates can be then understood as indicators of the text responses. Our Inbuilt Rubric method aims to transform the latent nature of vector space models into semantic spaces whose coordinates have a priori explicit semantic meanings, such as the meaning of the important concepts we want to identify and evaluate in texts. Basically, this method maps assessment rubrics into vector spaces that allow to obtain information about the presence or absence of the items of the rubric (here, semantic concepts). Two different versions of the method can be implemented, namely: using descriptors or nouns embedded in fragments of the instructional text. In this poster, we present a brief explanation of the computational method, an empirical illustration and how to implement it in R. Empirical results of convergent and discriminant validity support the use of these computational scores. In a broader perspective, this study defends the necessity of using a validity-centered approach to gather evidence in favor of computational scores to measure constructs from written materials. This poster summarizes and extends a recent publication: https://doi.org/10.3758/s13428-021-01764-6.

WEDNESDAY 3 JULY
Session 1
Topic: Artificial Intelligence in testing, psychological assessment and survey research

544. An Examination of Racial Bias in Artificial Intelligence When Conducting Neurological and Learning Assessments.

Joseph Kush
Duquesne University
Inna Vaisleib
University of Pittsburgh Medical Center

This proposed poster investigates the existence of racial bias embedded in the algorithms that underlie the popular Artificial Intelligence (AI) application, ChatGPT, when used to assist with the assessment of neurological and learning assessments in children. A series of 12 sets of vignettes were created describing common neurological conditions associated with learning disorders in children. Each of the sets contained identical language (e.g., age, medical history, etc.), apart from the racial background of the child (e.g., White, Black, Asian-American, Native American) which was intentionally, systematically varied. As anticipated, most of the diagnoses or recommendations produced by the AI-generated output were nearly identical, regardless of the racial background of the child. However, when disparities did occur, they were at times subtle while in other instances dramatic. The study accentuates the critical importance of recognizing and mitigating racial biases embedded within AI systems, when such technologies are employed with neurological and neuropsychological assessments. The poster offers cautions for practitioners, when utilizing AI-driven tools for neurological and learning-problem evaluations with cross- cultural or multi-cultural populations, and provides a novel, yet powerful methodology for future research examining potential biases in AI systems.

WEDNESDAY 3 JULY
Session 1
Topic: International assessment

617. Measurement invariance of a general cognitive performance measure across 26 European countries and Israel.

Adrián García Mollá, José Manuel Tomás, Amparo Oliver, Zaira Torres, Irene Fernández

Department of Methodology for the Behavioral Sciences, Faculty of Psychology University of Valencia, Spain

Background: The extended use of cognitive composite measures in scientific literature presents some shortcomings, such as reducing inter-domain variability by generating a global cognitive score or losing sensitivity for detecting subtle age-related changes. In short, research tends to employ these summary measures as markers of cognitive status. Objective: To study the psychometric properties of a popular measure of general cognitive performance and test cross-country measurement invariance. Methodology: Data for this study come from the 8th wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). The sample was composed of 46733 adults from 26 European countries and Israel with an average age of 70.27 years old (SD = 9.47). Over half of the sample was female (57.5%). Most of the participants either reported a level of upper secondary education or first stage of tertiary education. First, a Confirmatory Factor Analysis (CFA) of the one-dimensional structure of general cognitive performance was tested in the whole sample. Then, we examined the approximate measurement invariance of the measure using alignment. Results: Model fit was very satisfactory for the whole sample: 2(2) = 1540.23, p < .001, CFI = 0.957, SRMR = .040, RMSEA = .128 [90% CI: .123-.134]. In turn, alignment results showed substantial non invariance of item intercepts and limited invariance of items' factor loadings. Implications: Although results showed adequate functioning of the measure in the general sample, measurement invariance across countries does not hold. Therefore, researchers using SHARE data are advised against employing a composite measure of general cognitive performance, as simple comparison across groups is not possible. Adrián García Mollá is the recipient of grant PRE2022-102085 funded by MCIN/AEI/ 10.13039/501100011033 and by "ESF Investing in your future".

WEDNESDAY 3 JULY
Session 1
Topic: International assessment

550. Quantitative Literacy proficiency in Mathematics students: a cross-domain diagnostic exploration.

Tatiana Sango, Sanet Steyn
University of Cape Town

Cross-domain mapping in diagnostic assessment enhances performance data interpretation accuracy and reveals curricular associations. In South Africa, university readiness is assessed by National Benchmark Testing (NBT) instruments: Academic and Quantitative Literacy (AQL) and Mathematics (MAT). IRT-based scores categorise candidates into benchmarks (Proficient, Intermediate Upper/Lower, Basic), indicating needed support in university programmes (Degree/Diploma). NBT data serves diagnostic purposes, offering evidence-based insights into subject-specific and domain skills. This study focuses on the relationship between Mathematics and Quantitative Literacy, examining the performance of students who have taken Mathematics (as opposed to Mathematical Literacy as a school subject) in a Quantitative Literacy (QL) test. Typically, students with a Mathematics background exhibit higher mathematical competence compared to those with Mathematical Literacy, which only requires elementary mathematical and statistical knowledge. However, having mathematical proficiency doesn't automatically ensure quantitative literacy proficiency, as QL skills need explicit teaching to Mathematics students. Applying Cognitive Diagnostic Modelling (CDM) principles, we analyse within-domain and cross-domain associations for 1,050 students using NBT QL and MAT test instruments. Statistical analyses within each benchmark investigate how combinations of attributes impact the interpretation of ability. This work contributes to academic profile literature, examining students' competence across domains and discipline-specific skill sets.

WEDNESDAY 3 JULY
Session 1
Topic: Psychometric modeling

238. RaterLynx: A Shiny App for Incomplete Rating Designs of Rater-mediated Assessments.

Angel Arias Carleton
University / Canada

This work introduces RaterLynx, an open-access Shiny app developed to streamline the creation of connected incomplete rating designs for performance assessments. It is tailored for measures scored within the framework of many-facet Rasch measurement (MFRM; Linacre, 1989) and generates systematic links designs (see Wind & Jones, 2019). This tool stands as a valuable resource for researchers, local schools, and small testing firms, providing a user-friendly platform to generate rating designs that establish linkages between raters and performances. In turn, this yields a data collection plan free of data subsets, which is an important requirement for many-facet Rasch modelling. Traditionally, the evaluation of performance assessments involves scoring procedures that employ two independent raters who are encouraged to engage in discussions when score discrepancies arise. If consensus is not reached, an experienced senior rater acts as an adjudicator for the final score. The Alpha coefficient (Saxton et al., 2012) and Cohen's Kappa (Smit & Birri, 2014) are commonly used metrics for interrater reliability, but these metrics do not consider raters' severity, centrality, or leniency, leading to score quality problems. Conventional adjudication also falls short in addressing rater severity/leniency issues, as rater effects can also be inherent characteristics of the adjudicator. Many-facet Rasch measurement has gained popularity in performance assessment because it is a measurement model that provides valuable information on rater effects (e.g., severity, centrality, and fit statistics). However, a connected rating design is necessary before the data are analyzed. Raters must be linked through similar performances with other raters; otherwise, the rating design may contain subsets of data, which may yield incomplete or misleading information about rater effects. RaterLynx provides straightforward steps to generate connected rating designs for scoring performance assessments

WEDNESDAY 3 JULY
Session 1
Topic: Psychometric modeling

461. Comparative Analysis of Psychometric Models for Testlet-Structured Assessments.

Carlos David Diaz Lopez, Joaquín Caso Niebla
Universidad Autónoma de Baja California, México
Coral González Barbera
Universidad Complutense de Madrid, España

The objective of this research was to compare prevalent psychometric models applied to testlet-based tests. The comparison considered fit indices, item parameter magnitude and precision, score precision and fairness, and their effect on examinee classification. Through a systematic literature review, 10 psychometric models were selected from item response theory, testlet response theory, and the Bi-factor approach. These models were applied to a dataset of a reading comprehension test used for student selection (n = 30,853) in a Mexican public university. Model fit was assessed using goodness-of-fit indices and information indices (AIC, BIC, and DIC). Additionally, parameter magnitude and precision, as well as the reliability of scores generated by each model, were compared. To examine the effect of testlets on test fairness, differential item functioning detection was conducted through logistic regression. Finally, the capacity of each of the 10 models to select examinees based on their skill level was evaluated. Based on goodness-of-fit indices, the two-parameter testlet model (2P-TRT) emerged as the best-fitting model, followed by the two-parameter Bi-factor model (2PL-BM). Analysis of parameter magnitude and precision revealed that unidimensional models (I) underestimated the difficulty parameter, (II) overestimated the discrimination parameter of items, and (III) overestimated the precision of test takers' ability parameters, directly impacting the interpretation of test reliability. In summary, across all comparisons, models from the testlet response theory provided the most effective alternative for analyzing testlet-based tests. The evidence suggests that these models allow for control over the effects of item format on the reliability, comparability, fairness, and validity of intended interpretations for scores in the reading comprehension test.

WEDNESDAY 3 JULY
Session 1
Topic: Innovations in test development

193. Research & Invention Methodology to generate a Cross-Cultural Test.

Claudia Gusso
International Test Commission (ITC)

A standard model was designed and implemented by a female developer of a test that measures various aspects of the dynamic connection between a person's psychic and physical expressions in a personal environment. The test was developed using a synoptic pattern inspired by the human model as a reference system for measuring the variations in states and the stability of personality traits. In the field of psychological testing, the challenge focuses on a Research and Invention Methodology (R&IM) that offers facilitation for cross-cultural intervention; thus, to share innovation solutions the female inventor explains the utility of a new personality test: the Static Dynamic Personality Test (TPSD). A long-mixed validation process is currently underway with the involvement of experts, who address and evaluate original solutions of psychometric measure. Research is advancing on the production of a large quantity and quality of data collected from tests and conveyed into different cultural containers. In fact, to facilitate the translation and adaptation process, the various thematic categories in support of the interpretation of test results were first distinguished in boxes to sort out the vocabulary of typical concepts for each culture. As a result of the Research & Invention Methodology, the Static Dynamic Personality Test is patented in Italy, in the USA, and in the PRC. In this tri-cultural context, we will analyse examples of translation and adaptation through European (ITA), Asian (PRC), and American (USA) samples. A heated debate raises considerations of moral recognition regarding cultural differences in convergence and divergence, which emerge during the translation and adaptation procedures of the new personality test. CLAUDIA GUSSO (0000-0003-0489-0963) – ORCID

WEDNESDAY 3 JULY
Session 1
Topic: Innovations in test development

807. Stability Analysis of Estimation of Piecewise Linear ICCs.

Gen Hori
Asia University
Sayaka Arai
National Center for University Entrance Examinations / Japan

Although classical test theory and item response theory each have their own advantages and disadvantages, the former is still often used in actual educational settings because the latter requires a large number of subjects and specialized software. In previous studies, the authors have proposed a modified way to draw piecewise linear ICCs that reduces the drawbacks of the former, which are subject dependence and item dependence. In this study, the authors clarify through numerical experiments how the stability of the estimation of the piecewise linear ICC varies with the number of subjects.

WEDNESDAY 3 JULY
Session 1
Topic: Innovations in test development

539. Construction of a Situational Judgment Test for the selection of prison officer candidates: from job analysis to item production.

Gucek Richard, Loarer Even, Terriot Katia
Conservatoire National des arts et métiers

The aim of this paper is to present the stages in the construction of a Situational Judgment Test (SJT) to assess the decision-making skills of prison officers in France. These tests have the dual advantage of being closer to the behaviours used in contextualised situations, which allows a better assessment of skills than is possible with self-report questionnaires (see Lievens & al., 2008), while also allowing a degree of psychometric standardisation and easier administration, as they are easier to use than that of role-playing exercises or real-life situations (Lothe & al., 2012). However, the construction of these tools requires a precise analysis of the activity and of the conditions under which it is carried out, as well as a modelling of the processes involved in order to be able to assess professional performance (McDaniel et al. 2001, 2007). To analyze work activity, we carried out observations and semi-structured interviews with prison officers and individual interviews with prisoners. We also used the Fleishman Job Analysis Survey (FJAS) method (see Caughron & Mumford, 2012) to accurately identify the skills considered the most important in the work of prison officers. Content and lexicometric analyzes resulted in a model with 10 criteria: 4 situation-related criteria (urgency, criticality, climate and complexity) and 6 prisoner-related criteria (dangerousness, vulnerability, agency, compliance, openness and personal motivation). These criteria were cross-referenced with 6 pre-defined methods of action, which can be grouped into 3 categories: 1/direct (involving physical action or verbal orders); 2/ semi-directive (involving persuasion or negotiation skills); 3/ non-directive (involving participative or delegative options). Finally, this model was used to construct 12 items, each consisting of a "problem situation" scenario and a choice of 7 possible responses corresponding to the behavioural scenarios predicted by the model.

WEDNESDAY 3 JULY
Session 1
Topic: Innovations in test development

623. Relationship between interests differentiation and social-emotional skills in Brazilian students: A new interest differentiation index.

Gustavo Henrique Martins
Ayrton Senna Institute (Brazil) and Universidade São Francisco (Brazil)
Filip De Fruyt
Ghent University (Belgium), eduLab21 - Ayrton Senna Institute (Brazil), and Ayrton Senna Institute Chair@ Ghent University (Belgium)
Ana Carla Crispim
eduLab21 - Ayrton Senna Institute (Brazil)
Joyce Scheirlinckx
Ghent University (Belgium) and Ayrton Senna Institute Chair@Ghent University (Belgium)

This study examines how student vocational interests and differentiation of their interests are associated with social-emotional skills. Central questions are examining whether indicating a preference for an interest domain also implies that students perceive that they have well-developed skills associated with that domain, and whether such association is stronger for students with a differentiated single-interest profile, also called the 'interest-skill-differentiation' hypothesis. To test this hypothesis, we propose to explore differentiation of a particular RIASEC domain relative to all the other domains, by calculating the difference between the preference score for that domain minus the average for all other domains. Brazilian students in 5th or 9th grade in middle or 3rd year of high school (Total N= 234.857) were administered 18REST-2 (Martins et al., 2023) and SENNA (Primi et al., 2021) to describe vocational interests and social-emotional skills. Interests were distinctively associated with students' social-emotional skills, though associations were small to moderate. Differentiation of interests increased across adolescence especially for girls. Students reporting better developed skills had more differentiated interest patterns. Those with single-differentiated interests did not report higher developed social-emotional skills, with few exceptions, refuting the interest-skill-differentiation hypothesis. The present work underscored that vocational interests and social-emotional skills are distinct and only modestly related constructs. They are unique components in formative and summative assessment that should be considered in tandem to guide educational and vocational counseling.

WEDNESDAY 3 JULY
Session 1
Topic: Innovations in test development

806. A Framework for Detecting Both Main Effect and Interactive DIF in Multidimensional Forced-choice Assessments.

Kai Liu Liu
Jiangxi Normal University
Yi Zheng Zheng
Arizona State University
Siwei Peng Peng, Zhichen Guo Guo, Junhuan Wei Wei, Fangbin Chen Chen, Yan Cai Cai, Dongbo Tu Tu
Jiangxi Normal University

In recent decades, multidimensional forced-choice (MFC) tests have gained widespread popularity in organizational settings due to their effectiveness in reducing response biases. Detecting differential item functioning (DIF) is crucial in developing MFC tests, as it directly impacts test fairness and validity. However, existing methods appear insufficient for detecting DIF induced by the interaction between multiple covariates. Furthermore, for multi-category, ordered and continuous covariates without protected classes prescribed by laws or regulations, existing approaches often dichotomize them using a-priori cutoffs, commonly using the median of the covariates. This may lead to information loss and reduced power in detecting MFC DIF. To address these limitations, we propose a method to identify both main effect DIF and interactive DIF. This method can automatically search for the optimal cutoffs for ordered and continuous covariates without pre-defined cutoffs. We introduce the rationale behind the proposed method and evaluate its performance through three Monte Carlo simulation studies. Results demonstrate that the proposed method effectively identifies various DIF forms in MFC tests, thereby increasing detection power. Finally, we provide an empirical application to illustrate its practical applicability.

WEDNESDAY 3 JULY
Session 1
Topic: Innovations in test development

485. Advanced Care Planning in Mental Health: From Systematic Review to Instrument Development.

Maite Barrios, Chao Zhang, Ángela Berrío, Juana Gómez-Benito, Georgina Guilera
University of Barcelona/Spain

Advanced Care Planning (ACP) is increasingly recognized as a critical component in medical care, emphasizing the autonomy and dignity of individuals across various health conditions. This is particularly pertinent in mental health, where ACP enables adults with mental illnesses to articulate their future healthcare preferences and decisions, ensuring respect for their choices during crises when their decision-making capacity might be impaired. However, there is a notable lack of instruments tailored to mental health. Our study describes the methodology adopted in designing specific instruments for evaluating ACP-related aspects in mental health. Initially, we conducted a systematic review of existing studies that developed or adapted measurement tools for ACP assessment in diverse health conditions. This review encompassed a comprehensive literature search across five databases: PsycNET, PubMED, Web of Science, Scopus, and CINAHL. The aim was to identify studies centered on the development, adaptation, and validation of ACP assessment tools. Subsequently, we identified relevant aspects measured by these instruments and the items that could be integrated into a new tool tailored for ACP in mental health. Out of 1,222 studies, 54 met our inclusion criteria, contributing to a total of 57 instruments. The majority of these instruments evaluated knowledge of ACP processes (36%), attitudes towards ACP (17%), engagement in ACP (16%), and self-efficacy in ACP (12%). We reviewed and adapted items focusing on knowledge and attitudes for use in the mental health context. This procedure led to the development of two novel instruments specifically designed to assess attitudes and knowledge about ACP in mental health. These instruments allow us to address previously under-evaluated aspects in the ACP process in mental health, potentially fostering the promotion of a culture of empowerment and enhancing the quality of care.

WEDNESDAY 3 JULY
Session 1
Topic: Innovations in test development

793. Student Evaluation in a Diverse Educational Environment: India's Holistic Progress Card.

Paul Borysewicz
ETS/USA
Aakanksha Bhatia, Neeraj Venkatamaran
ETS/India
Jonas Bertling
ETS/USA
Indrani Bhaduri
PARAKH/India

Educators have long recognized the importance of encouraging and measuring learners' development in non-cognitive domains. Implementing such programs has proven challenging even in culturally homogenous countries with high levels of economic development. This presentation will provide a review of a new initiative by India's Ministry of Education to implement holistic pedagogy and evaluation in India's highly diverse educational environment. The Holistic Progress Card (HPC), which is being piloted this year, shifts focus to learners' cognitive, social/emotional, and creative development in integrated activities and classroom interactions. Drawing on evidence-centered design, multiple intelligence theory, student-centered learning, project-based learning, and problem-centered learning, the activities aim to provide teachers with workable models both for developing and evaluating learners' performance. India's linguistic diversity, the varied experience level of its teachers, the decentralized structure of its educational system, the continuing influence of traditional rote-learning, as well variations of available resources all pose extraordinary challenges to implementing a program like HPC. This presentation will discuss several strategies taken to mitigate those challenges. Among these mitigation strategies was the decision to construct model, modular rubrics for evaluating student performance across different abilities in the same activity. Rubrics are also intended to be shared with the learners to facilitate their participation in their own development. In all aspects of the HPC, a high degree of flexibility and adaptability as per the needs of the learners is encouraged. Small-scale piloting of the HPC has been supplemented with feedback questionnaires currently being analyzed. Significant quantitative or qualitative findings from the questionnaires will be included in the presentation as appropriate.

WEDNESDAY 3 JULY
Session 1
Topic: Innovations in test development

618. Towards the Development of a Grade 4 Science Achievement Test.

Pia Marie Therese Papa, Mary Nela Aguila
Global Resources for Assessment Curriculum and Evaluation, Inc.

In 2012, the K-12 Basic Education Curriculum Framework was constituted by the Department of Education (DepEd) to develop Filipinos' readiness and mastery for the 21st Century world of work (DepEd, 2015). To further aid educators and guide students in measuring their proficiency in the core subjects of Philippine education, an achievement test for Grade 4 Science was designed and developed. The researchers opted to develop this because the DepEd curriculum aims to ensure that it reflects the different developmental milestones of this stage and to align it with the goals of the TIMSS international large-scale assessment. In the process of test development of the Grade 4 Science achievement test, it underwent a series of writing and review by subject matter experts in Science to establish its content validity. It also utilized the item response theory analysis using a 3-parameter logistic model (3PL), which uses discrimination, difficulty, and guessing as parameters. One hundred items for Grade 4 Science were developed and administered to students of 24 schools across Luzon, Visayas, and Mindanao islands of the Philippines for pilot testing. The initial development resulted in a 50-item test. Recommendations for further validity studies are provided in this research.

WEDNESDAY 3 JULY
Session 1
Topic: Innovations in test development

459. Damage to Honor Scale: A Forensic Psychological Contribution to the Evaluation of Moral Damage.

Alejandra del Carmen Domínguez-Espinosa
Universidad Iberoamericana, México
Yessica Daniela González Berriel2/2, Marina Flores-Camargo
Universidad Nacional Autónoma de México, México

The spread of false information about something that we "allegedly" said or did is a situation that people are susceptible to experiencing. Esbec (2000; in Muñoz, 2013) and Echeburúa et al. (2004), who defend that moral and psychological damage are different concepts, reject to assess the damage to honor, considering it subjective and, therefore, not subject to measurement. Despite this, the present study aimed to contribute a conceptualization and operationalization, through a self-report scale, damage to honor as part of moral damage in the field of forensic psychology. The present study was non-experimental, correlational, exploratory and developed on field. A total of 674 people participated. The "Jamovi" statistical program was used. This study considered four phases for its psychometric development: 1) content evidence (the original items were developed and 2 versions of the scale were obtained); 2) evidence of internal structure (for the first version, 4 factors with 23 items were obtained); and, for the second, 5 factors with 32 items were obtained) and precision indices (above = .85 and = .74 for each version); 3) analysis of parallel forms (with correlations above r= .75); and, 4) evidence based on relationships with other variables (obtaining a relationship with the variables of emotional-negative coping style, external locus control and social support). In summary, the two versions of the Damage to Honor Scale were obtained, both of them have acceptable accuracy rates and evidence of validity, therefore, they can be used as part of the process of measuring damage to honor to identify whether a person has suffered such damage. Finally, this study represents a necessary contribution to the practice of the forensic psychologists.

WEDNESDAY 3 JULY
Session 2.1
Topic: Innovations in test development

258. Interactions Between Big Five and Dark Side Traits Utilising Factor Analysis of Personality Assessment Data

Mikael Nederström, Anita Rintala-Rasmus
Psycon/Finland

In our study we aimed at a deeper understanding of assessing universal personality-based traits in W/O setting. The study employed a comprehensive joint factor analysis to explore the interplay between Big Five personality traits and dark side traits within the framework of W/O psychology and assessment. Drawing on a large sample of real-life personnel assessment participants, N=9225, we used joint factor analysis to investigate the shared and unique variance between the traditionally established B5 traits and the less-explored dark side traits, including the 11 maladaptive traits from the DSM-V framework. The methods included the SRS test measuring 11 personality disorder traits (Nederström & Furnham, 2012) and a B5-based test (Nederström, 2023), where each of the five global traits is measured by two separate facets. Our findings demonstrated patterns of co-variation, shedding light on the relationships between adaptive and maladaptive personality traits. Notably, certain facets within the B5 exhibit associations with specific dark side traits, contributing to a more nuanced understanding of personality dynamics. The dataset formed four B5 factors, with Openness to experience notably absent. Extraversion loaded most strongly on narcissistic & histrionic traits, Neuroticism on paranoid & schizotypal traits, Conscientiousness on obsessive traits, and Agreeableness inversely on psychopathic & schizoid traits. This study advances our theoretical understanding of personality and has practical implications for assessing the normal and maladaptive trait continuum in W/O psychology. By integrating B5 and dark side trait frameworks, our findings contribute to a more holistic perspective on personality, emphasizing the importance of considering both adaptive and maladaptive traits in assessment. Recruitment decisions based on valid assessment support wellbeing and succeeding in the worklife and produce positive candidate experience in recruiting from the global talent pool.

WEDNESDAY 3 JULY
Session 2.1
Topic: Innovations in test development

270. A Proposal and Comparison of Item Selection Methods in Computerized Adaptive Testing for Multidimensional Forced-Choice Measures

Qin Wang
China
Yi Zheng
USA
Yan Cai, Dongbo Tu, Daxun Wang, Fen Luo, Kai Liu, Junhuan Wei, Zhichen Guo, Fangbin Chen, Jiyuan Ding, Xuhong Song, Pan Jiang
China

Computerized adaptive testing for multidimensional forced-choice measures (MFC-CAT) integrate the CAT technology into the MFC measures that leverages the strengths of both. As a crucial component of MFC-CAT, the existing item selection methods of MFC-CAT can be divided into two types—item selection methods based on Fisher information (FI-based methods) and Kullback-Leibler information (KL-based methods). FI-based methods have been criticized for the attenuation paradox issue and over exposure. In contrast, KL-based methods offer improved estimation accuracy but provide no significant advantage in controlling item exposure rates and ensuring even usage of the item pool. Furthermore, the impetus of existing study was the lack of guidance in the literature of MFC-CAT in terms of which item selection method to use and under what conditions. Therefore, this study proposed a new type of item selection methods based on the Bayesian theory (Bayes-based methods) to improve the uniformity of item pool usage without sacrificing the estimation accuracy, and make a comprehensive comparison with the two existing types of methods. We conducted three simulation studies in three-dimensional scenario, seven-dimensional scenario and real data scenario. Our findings showed that three types of item selection methods can be applied to MFC-CAT. FI-based methods excel in efficiency, whereas KL-based MFC-KB method and Bayes-based MFC-CEM method are preferable for estimation accuracy. For more uniform item pool usage, the new Bayes-based methods are recommended. The study concludes with a table of method recommendations for various conditions to assist researchers and practitioners in MFC-CAT application and development. This study is expected to guide future item selection and applications of MFC-CATs.

WEDNESDAY 3 JULY
Session 2.1
Topic: Innovations in test development

300. Playful testing of executive functions with Yellow-Red: Tablet-based battery for children between 6 and 12 available in 8 different languages

Ricardo Rosas P.
Universidad Católica de Chile

Executive Functions are psychological processes of great importance for the proper functioning in various areas of human development, including academic performance. For this reason, from both clinical and educational perspectives, there is great interest in how they are assessed. This presentation will show Yellow-Red, an instrument for directly assessing executive functions in children between 6 and 11 years of age, in a playful format using a digital support. The test is based on a 3-factor model of executive functioning: Inhibition, Working Memory, and Cognitive Flexibility. Yellow-Red comprises six subtests: cognitive inhibition (2), behavioral inhibition (1), working memory (2), and cognitive flexibility (1). The presentation will show the test's theoretical foundations, the standardization and validation methodology, and the factorial results in detail. It will also show how to interpret individual reports and learn how to access the tool free of charge for research purposes. The test is standardised for Chile, is in the process of being standardised for Germany, and has versions in Spanish, German, Norwegian, Hungarian, Romanian, Chinese, Thailand, and English. Finally, comparative results between all these countries will be shown.

WEDNESDAY 3 JULY
Session 2.1
Topic: Innovations in test development

302. Distractor Analysis of Eye Movements for Multiple-Choice Questions

Ergun Cihat Corbaci,
Dr. / Turkiye
Nilufer Kahraman
Prof. Dr.

In educational assessment, the significance of understanding how students engage with multiple-choice questions (MCQs) cannot be overstated. The correct option and the distractors (the incorrect options) in MCQs play a pivotal role not just in assessing knowledge but also in understanding the cognitive aspects of item processing and decision-making. This research uses real-time eye fixations to investigate distractor gaze patterns of students while they were navigating through MCQs. Application data were seventy-one examinees' responses to a five-multiple-choice item test experiment (items 1 to 5, having fixed correct-coded response placements of A to E, respectively) measuring reading comprehension skills in English. A number of choice-fixation metrics, such as choice fixation durations and revisit counts, were computed for each item. Results show that the eye fixation patterns of the examinees can provide unbiased, clear, and easy-to-interpret information about how each distractor functions relative to the others and the correct coded choice. The logistic regression analysis showed that the examinee response (correct versus incorrect) and the placement location of the correct-coded choice were meaningful predictors of the choice-fixation patterns. Besides, it was found that some distractors, regardless of their placement order, had more revisits from one of the response groups, and the differences between revisit means were significant according to ANOVA results. The results are promising and suggest that distractor analysis of eye fixation metrics can be instrumental when combined with the already available examinee and item indicators, such as examinee average item pacing times (in Sec.) and item difficulties. The outcomes of this research hold significant implications for psychometricians, assessment developers, and educational researchers seeking to enhance the validity and reliability of multiple-choice assessment tools. This study was partially supported by Gazi School of Education and by TUBITAK under grant SOBAG 120K142. Keywords: Eye-tracking, Distractor,

WEDNESDAY 3 JULY
Session 2.1
Topic: Innovations in test development

385. **Innovations in Skills-based Assessment**

Lydia Liu
ETS

The landscape of modern workforce is rapidly changing due to technological advancements, which requires students and workers to demonstrate durable skills to be fast learners, productive collaborators, and effective communicators, to be future proof. Educators, employers, and policymakers ask for skills-based learning and hiring, yet the efficacy of such efforts is often hindered by the lack of effective tools for accurately quantifying skills. Two specific challenges stand out when it comes to skills assessment. One is the complex nature of many durable skills such as critical thinking and collaborative problem solving, which are multidimensional and require delicate balance between sufficient construct representation and psychometric practicality. The other challenge concerns capturing the broad sources of skills acquisition—skills don't just come from formal educational activities but from many experiences outside of school. This paper illustrates our approach in defining skills drawing on a wide range of literature and frameworks designed to identify skills vital for the future workforce. Our approach involves a thorough analysis of international skills frameworks such as the World Economic Forum, OECD, O'NET in the U.S., India Skills Report, and the UAE. We delve into the rich affordances of identified skills and the indicators that can be used to quantify the skills. For example, a learner can demonstrate perseverance through excelling in athletics, pursuing a musical journey, taking care of younger siblings, or walking to school for an hour every morning and be on time because their family can't afford the school bus. There are many different ways learners can demonstrate their skills. Having a comprehensive system to capture and quantify the skills so that learners, especially ones from under-served background, have a skills record that they can carry with them into future academic and workforce settings has great potential to promote economic mobility.

WEDNESDAY 3 JULY
Session 2.2 SYMPOSIUM
Topic: International assessment

509. Revising and enhancing the EFPA Test Review Model with lessons learned in practice

Nigel Evans
NEC
Discussant name
Discussant surname
Discussant affiliation

The European Federation of Psychologists' Associations (EFPA) Board of Assessment (BoA) is updating their Test Review Model (TRM). The latest version of the EFPA TRM was published in 2013. This review model is extensively used as a standard across Europe in evaluating tests (with many translations around the world). Along with the utilization of technology in assessment, new topics have arisen since the inception of the latest version of the EFPA TRM so a revision is necessary to remain up-to-date. This symposium looks at the progress and challenges of revising the TRM, with a review of the work of the BoA TRM task force to date. Most necessary changes appear as additions reflecting advances in testing and psychometrics, including continuous norming and random sampling; new views on reliability and validity issues, local vs global norming, the use of hidden algorithms, variations of online/digital testing, gamification, and the introduction of Artificial Intelligence into many aspects of test development. Illustrations of using the model in three different countries follow to highlight the breadth and depth of the model's application, with lessons learned feeding into the TRM review process. All are valuable to help achieve the main goal of the EFPA Test Review Model - providing a description and a detailed and rigorous assessment of the tests, scales and questionnaires used in the field of psychological and educational assessment

WEDNESDAY 3 JULY
Session 2.2 SYMPOSIUM
Topic: International assessment

568. Implementation of the Test Review Model in Spain: Impact, Improvements and Challenges (Translation of tests, psychological assessment instruments and survey questionnaire)

Ana Hernández
University of Valencia
Paula Elosua
University of the Basque country
Francisco J. Abad
Autonomous University of Madrid
José R. Fernández-Hermida
Spanish Association of Psychology
José Muñiz
Nebrija University

The Spanish Test Review Model (TRM) designed to evaluate the quality of tests, was originally proposed in 2000 (Prieto & Muñiz, 2000), and first implemented in 2011 (Muñiz et al., 2011). Since then, the Spanish Test Commission has evaluated over 100 tests, with the results made available open access to the community. Over the years, the model and its implementation process have undergone various improvements (Abad, 2024; Hernández et al., 2006, 2015). This presentation outlines these key enhancements and assesses their impact on test quality. We particularly focus on comparing tests reviewed before and after the introduction of key TRM changes, with special attention to updated editions that replaced previously reviewed versions. Our findings show that test ratings have remained fairly consistent across most areas, with notable improvements in Differential Item Functioning assessment, Item Response Theory application, and inter-rater reliability. Comparisons between older and newer test versions indicate variable scores, with increases in some areas and decreases in others. This variability suggests that while modifications in the TRM and its application have raised standards to some extent, certain degree of subjectivity in applying TRM criteria still exists. Overall, the TRM's impact on enhancing test quality appears positive, but challenges such as inconsistent application of standards and the balancing of local versus international studies remain. This, together with the increasing impact of new technological developments in testing, underscores the importance of ongoing updates to address evolving challenges in psychological and educational testing.

WEDNESDAY 3 JULY
Session 2.2 SYMPOSIUM
Topic: International assessment

570. Cross-Cultural Testing – recent observations from the British Psychological Society Test Review Process (Validity and fairness in cross-cultural testing, psychological assessment and survey research)

Charlie Eyre
Consultant Editor, British Psychological Society Test Reviews, member of BPS Committee for Test Standards. Director, Workspheres Ltd.
Iain Coyne
Senior Editor, British Psychological Society Test Reviews, member of BPS Committee for Test Standards. Reader in Organisational Psychology, Loughborough Business School
Nigel Evans
Director NEC, UK)

The opportunity for psychometrics to be used across international markets has rapidly increased through advances in online testing capability and the increased globalisation of business. This has led to a proliferation of tests available to users across a broad range of global markets. However, varying practices exist in terms of the local adaptation and norming of tests for different national and cultural contexts. The BPS Test Review process, like other European test review systems, is underpinned by the EFPA Test Review criteria (EFPA, 2013). This presentation reviews the broad trends in recent years for the BPS Test Review process, drawing evidence from a broad sample of published test reviews. In particular, it reflects on the factors influencing the evaluation of the test's local adaptation and norming, drawing the EFPA Test Review criteria relating to norms (7.2.2.2, 7.2.4, Section 9), as well as the ITC Guidelines for Translating and Adapting Tests (ITC, 2017). The review will draw out themes around the development and use of global norm groups, the interpretation guidance being provided to test users by publishers, observed differences in sampling strategies, and the extent to which score equivalence is being evidenced across different cultures and regions. Conclusions will be drawn on the opportunities and challenges for test publishers in developing effective testing solutions across differing cultural and national contexts. Consideration will also be given to the specific challenges of applying standardised test review criteria in an increasingly diverse, cross-cultural test publication market.

WEDNESDAY 3 JULY
Session 2.2 SYMPOSIUM
Topic: International assessment

532. Two applications of the EFPA Test Review Model in Norway (Translation of tests, psychological assessment instruments and survey questionnaire)

Siri S. Helland
RBUP, Pilar, Norway
Rudi Myrvang
Oslo Hospital, Norway
Brynhildur Axelsdottir, Simon Peter Neumer
RBUP, Pilar, Norway
Jannike Kaasbøll, Kenneth Stensen
RKBU Mid, Norway
Kjell Morten Stormark
RKBU Vest; Norce, Norway
Monica Martinussen
RKBU Nord, UiT The Arctiv University of Norway

Tests are used for many purposes and in different contexts. This includes screening and assessment in clinical and educational settings as well as personnel selection in civilian and military organizations. In many cases test results are used either alone or combined with other types of information to inform decisions. Norway includes approx. 5,5 mil. people, and many of the tests used are developed in other countries in a different language and cultural background. The purpose of this paper is to describe two different examples on how the EFPA test review form can be used to examine the quality of tests both for children and young people as well as for adults in general. Children and young people The test review process for tests appropriate for children and young people is organized as a collaboration between four different regional centers for child mental health funded by the Norwegian Directorate of Health. The work is organized as an open access journal which publishes systematic reviews of tests used in Norway. The paper will describe how the EFPA criteria are used to evaluate psychometric properties. So far, the journal (Psyktestbarn.no) has published 82 articles and the journal website was visited by 45 500 different users last year. Adults When it comes to test review process for tests used for adults, there is a well-established process for review based on the EFPA criteria within the field of organizational psychology in Norway. Where test publishers can have their instruments evaluated by DNV (https://www.dnv.com/services/design-verification-and-quality-assurance-2999). The process concludes in a written report and a certificate if the instrument is considered to meet the criteria (https://www.dnv.no/services/certification-of-occupational-test-tools--163112). This test review process could act as an inspiration for other fields of psychology where tests are used (clinical and educational settings), where there currently lacks quality control.

WEDNESDAY 3 JULY
Session 2.2 SYMPOSIUM
Topic: International assessment

527. The revision of the EFPA Board of Assessment Test Review Model: the last hurdles on the way to a necessary and thorough update (International assessment)

Schittekatte Mark1/1, Evans Nigel2/2
(1) Ghent University, Belgium, (2) Director NEC, UK

The main goal of the EFPA Test Review Model (TRM) is to provide a description and a detailed and rigorous assessment of the tests, scales and questionnaires mainly used in the field of psychological and educational assessment. In other words, offering a tool for assessing the quality of tests. This information is made available to test users and professionals, in order to improve tests and testing, and help them to make the right assessment decisions. The EFPA TRM is part of the information strategy of the EFPA, which aims to provide all necessary mainly technical information about the tests in order to enhance its use. This EFPA test review model aims further to support and encourage the process of harmonizing the quality standards and the reviewing of tests across Europe. However the latest version of this Test Review Model dates from 2013, so a revision is very necessary! The EFPA Board of Assessment (BoA) is in the final phase of updating the TRM in this period, and how this is handled, with who (different stakeholders) and in what time frame, are few of the issues to be discussed. Further attention is given to: what topics are/were the most urgent to update (e.g. online testing, gamification, AI algorithms) and how can this TRM be implemented by local test commissions. Also the hurdles in this process in different European countries will be highlighted (e.g. reviewed by who, whether or not an overall final score for each reviewed instrument, need for financial resources, and considering the impact of negative reviews).

WEDNESDAY 3 JULY
Session 2.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

430. Current Trends and Best Practices in Cross-Lingual Assessment

Louise Badham
International Baccalaureate & University of Oxford / UK
Stephen Sireci
University of Massachusetts Amherst / USA
Malena Oliveri
University of Nebraska–Lincoln & Buros Center for Testing/ USA
Steve Dept
cApStAn linguistic quality control / Belgium

An increasingly common feature of 21st-century educational assessment is assessing students who interact with the world using different languages. In these situations, multiple language versions of assessments are needed. Developing multiple language versions of an assessment requires several layers of qualitative and quantitative processes that may involve translating (adapting) assessment materials developed in one (source) language to one or more other (target) languages, parallel development of multiple language versions of an assessment, and statistical adjustments of scores from different language versions of assessments to promote or evaluate comparable interpretations. In this symposium, we bring together practitioners and researchers with extensive experience in cross-lingual assessment. The symposium comprises four presentations and a discussant. The first presentation reviews the literature on methods for developing and evaluating cross-lingual assessments. The second presentation discusses the process of adapting source language versions of tests into other languages. The third presentation discusses the development of different language versions of an exam in a parallel fashion, where test content is developed separately in each language according to a common blueprint. The fourth presentation describes the comprehensive adaptation, quality assurance and quality control procedures involved in international large-scale assessments. Ample time will be provided for discussion with the audience.

Discussant name: Guillermo
Discussant surname: Solano-Flores
Discussant affiliation: Stanford University

WEDNESDAY 3 JULY
Session 2.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

435. Developing and Evaluating Exams for Use Across Multiple Languages: The Successive (Adaptation) Model (Translation of tests, psychological assessment instruments and survey questionnaire)

Maria Elena Oliveri
Buros Center for Testing
Stephen Sireci
University of Massachusetts
Louise Badham
International Baccalaureate Organization

In this presentation, we will delve into the intricate process of translating and adapting International Baccalaureate (IB) exams, focusing on the application of the successive adaptation model. The successive adaptation model serves as a robust framework employed by the IB program to adapt assessments across subjects such as Individuals and Societies, Science, Math, and Arts. The successive (adaptation) model involves the development of exams in a source language, typically English, followed by the translation (or adaptation) of these exams into other languages. The objective is to maintain the utmost similarity between the translated versions and the source language. Within the successive adaptation approach, a test is initially created in one language and subsequently adapted into target language(s) by one or more bilingual translators to ensure that the construct being measured remains consistent across languages while allowing for nuanced adjustments to align with the cultural context of the target languages. Thus, the successive adaptation process is not merely a process of linguistic translation; instead, it is a comprehensive strategy designed to preserve the integrity of the assessment while accommodating cultural subtleties to achieve cross-lingual measurement comparability. By employing this model, the IB aims to maintain the reliability and validity of its exams across diverse linguistic and cultural contexts, ensuring that the assessments accurately measure the intended constructs while being sensitive to the cultural nuances of the test-takers.

WEDNESDAY 3 JULY
Session 2.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

458. Developing, Facilitating, and Evaluating Validity and Comparability in Multi-Language Assessment Programs: A Review of the Literature (Translation of tests, psychological assessment instruments and survey questionnaire)

Stephen Sireci
University of Massachusetts Amherst, USA
Maria Elena Oliveri
University of Nebraska Lincoln
Louise Badham
International Baccalaureate Organization

Developing multiple language versions of an assessment often requires several layers of qualitative and quantitative processes ranging from translating (adapting) assessment materials in one (source) language to one or more other (target) languages, to statistical adjustments of scores from different language versions of the assessments to promote comparable score interpretations. Such "cross-lingual" assessment involves consideration of linguistic diversity from the earliest stages of test development through scoring and reporting of test results, as well as validation of the results. In this presentation, we review and summarize the literature with respect to ensuring validity in cross-lingual assessment programs. In particular, we focus on (a) test development activities that involve translation/adaptation using simultaneous or parallel development; (b) qualitative and statistical procedures for evaluating translated assessment material; (c) methods used to "link" or make more comparable scores from different language versions of assessments; and (d) methods for evaluating the results of cross-lingual assessments.

WEDNESDAY 3 JULY
Session 2.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

433. Best Practices in Adapting Test Across Languages: Lessons from cApStAn's Twenty-Five Years of Trial and Error (Translation of tests, psychological assessment instruments and survey questionnaire)

Steve Y. F. DEPT
cApStAn

In the late 1970s and early 1980s, comparative researchers established that variability in translation quality could compromise cross-country equivalence in multilingual data collection. Merely comparing back translations of the translated items to the initial version fails to detect sources of construct-irrelevant variance such as culture-driven perception shifts. More importantly, it does not consider that intentional deviations from the master version may be necessary to maintain equivalence. Reliance on bilingual reviewers is useful but requires a standardized approach. The International Test Commission's Guidelines for Translating and Adapting Tests (1996, 2002, 2005 for the First Edition; 2017 for the 2nd Edition) offered methodological ideas, many of which were implemented in the emerging field of international large-scale assessments (ILSAs) such as OECD/PISA, IEA/TIMSS, IEA/PIRLS, CONFEMEN/PASEC or UNICEF/SEA-PLM. The author of this paper supervised linguistic quality assurance of ILSAs for the past 25 years and shares empirical observations on the relevance and limitations of the criteria and metrics used to assess linguistic equivalence. In this presentation, we shall examine the prerequisites of a robust test adaptation design, the tension between the gold standard and budget constraints, attempts to reduce the inherent subjectivity of translation, and reporting practices. While progress has been made in terms of cross-linguistic equivalence of assessment instruments, there are limitations to diagnostics and corrective action suggested by both linguists and subject matter experts. Proxy indicators provide a measure of compliance with a set of guidelines. They do not account for all language/item interactions, nor do they provide a measure of functional equivalence. Measurement equivalence can only be examined by analyzing the data. Translated/adapted versions need to be piloted, e.g. by means of cognitive labs, online pre-tests or larger field trials.

WEDNESDAY 3 JULY
Session 2.5
Topic: Psychometric modeling

118. **Advances in testing MRI factorial models**

Agustin Martinez-Molina
Universidad Autonoma de Madrid
Elena Lacomba-Arnau
Universidad de Lleida
Luis Garrido
Pontificia Universidad Catolica Madre y Maestra
Patricia Rosell-Negre, Alfonso Barros-Loscertales
Universidad Jaume I

The fusion of factorial analysis with magnetic resonance (MR) data is a pioneering venture in scientific research. With few precedents, such as the work of Cooper et al. and van Kesteren, this field demands a dual expertise: advanced factorial techniques and sophisticated MR data handling. Researchers must navigate Exploratory and Confirmatory Factor Analysis, Structural Equation Modeling, and MR-specific protocols to extract and correct data. The intellectual challenge extends beyond technical prowess, requiring conceptualization and theoretical framing of results. This interdisciplinary realm necessitates collaborative teams well-versed in the mathematical and physical sciences. As we delve deeper into this niche, methodological precision is paramount. The correct choice between exploratory and confirmatory factorial approaches, and the accurate determination of dimensionality through methods like parallel analysis or EGA, are critical for maintaining the integrity of the research. Our study leverages a database of neural density from 300 subjects, aligned with Reinforcement Sensitivity Theory, to explore these complex methodologies. By parceling homotopic brain regions and employing factorial and network models (e.g., EFA, CFA, ESEM), we challenge the conventional three-factor structure of brain behavior interaction (BIS, BAS and Constraint), advocating for a nuanced understanding that traditional parceling and estimators remain relevant. The findings suggest a departure from established theories, with the best-fitting models demanding a reevaluation of potential method factors. This research not only advances our grasp of factorial and MR data integration but also reinforces the need for methodological exactness and interdisciplinary cooperation in the evolution of cognitive neuroscience.

WEDNESDAY 3 JULY
Session 2.5
Topic: Psychometric modeling

245. Modelling the relationship between problem-solving solution attainment and strategies using task-general process data with IRTrees

Huseyin Yildiz, Nathan Zoanetti
Australian Council for Educational Research/Australia

Theoretical Framework Besides whether students successfully reach the goal of interactive problem-solving tasks, they may also differ in the cognitive and self-regulatory strategies they adopt when solving them. Process data collected from event logs provide useful markers of individual differences in strategies. IRTrees (de Boeck and Partchev, 2012) are theoretically well-suited for modelling the relationship between different latent traits that may drive differences in strategy-related process data patterns and goal attainment. Objectives Our objective was to understand the nature and magnitude of the relationships between different latent variables specified to influence different aspects of problem-solving attainment and strategies. Methodology This study analyzed data from tasks designed to measure search-based problem solving (Zoanetti, 2010). Event logs were parsed to produce theoretically salient process data variables, including time-to-first-action, time-on-task, and the total number of student-task-interactions. The mentioned variables were used to apply several IRTree models. Results The relationship strength between process-focused and attainment-focused latent variables varied among models. In some cases, weak associations were found where theoretically stronger associations were expected. An example includes the latent correlation of 0.55 between the ability to solve problems and the propensity to solve them with relatively fewer interactions, suggesting that these latent variables tap into distinct traits. Implications These models provide a basis for testing scoring assumptions, for example determining whether process data and accuracy data can be combined into strictly ordinal score categories. In the example with a latent correlation of 0.55, adopting task-general partial credit scoring heuristics with one score point for an accurate solution and two score points for an accurate solution with fewer interactions would be questionable.

WEDNESDAY 3 JULY
Session 2.5
Topic: Psychometric modeling

474. Monk: Developing a vertical scale based on IRT to assess Math competence in primary education in the Spanish education system

Miguel A. Sorrel
Autonomous University of Madrid
David Cabello, Carolina Gamboa, Javier Pardo
Smartick

In response to Spain's challenging outcomes in international mathematical ability tests and the difficulties observed by teachers in classrooms, we present 'Monk', an innovative tool for adaptive assessment and learning. This tool includes a bank of more than 1,200 primary school problems (1st-6th grade) rigorously designed to understand the learning needs of each student at any given moment. It provides estimates of global mathematical competence, a breakdown of their strengths and weaknesses in each competency area, and the attitude towards mathematics and the risk of dyscalculia. Mathematical competence is calculated through a computerized adaptive test that takes about 20-30 minutes per test taker. In this presentation, we will focus on presenting the item bank calibration process and the vertical scaling followed. Specifically, the process included a project of item calibration under the Rasch model and the alignment of parameters using the Mean/Mean method, ensuring the creation of a common metric centered at the 3rd Primary level. The method followed and the results obtained in this respect will be presented. In addition, the evaluation tool formed from this item bank will be illustrated. To date, this platform has already been used by 50 centers across Spain, facilitating continuous monitoring of student progress by teachers, and allowing for more timely and effective pedagogical interventions. The evidence available to date allows us to state that, by specifically addressing individual student needs, Monk represents a significant step towards improving academic performance and mathematical understanding.

WEDNESDAY 3 JULY
Session 2.5
Topic: Psychometric modeling

500. Dimensionality Assessment in Forced Choice Questionnaires: First Steps Towards an Exploratory Framework

Diego F. Graña
Universidad Autónoma de Madrid
Rodrigo S. Kreitchmann
Universidad Nacional de Educación a Distancia
Miguel A. Sorrel
Universidad Autónoma de Madrid
Luis E. Garrido
Pontificia Universidad Católica Madre y Maestra
Francisco J. Abad5/5
Universidad Autónoma de Madrid

The popularity of forced-choice (FC) questionnaires has increased in recent years due to their ability to account for certain relevant and well known response style biases like social desirability. However, their design still entails certain difficulties. The dimensionality of FC data largely depends on the characteristics of the stimuli assembled into blocks. Additionally, within the widely employed confirmatory framework, single-stimulus data structure is assumed to be perfectly generalizable to the stimuli in FC data, which may not align with empirical data. As a first step to a more exploratory framework, the main goal of the present study is to determine if typical dimensionality assessment methods such as the Kaiser rule, empirical Kaiser, Parallel analysis, and Exploratory Graph Analysis are viable with FC data. To this aim, we conducted a Monte Carlo simulation study generating data from the Thurstonian IRT model. We manipulated as factors: the inclusion of unequally keyed items; the number of factors, variables and response options (generating graded preference conditions); the loadings mean and range; the factor correlations size, and the sample size. Then, we analyzed the hit rate and bias of the different dimensionality assessment methods recovering the true dimensionality of the data. Results highlighted Parallel analysis as the best method to assess dimensionality in FC questionnaires, suggesting that dimensionality assessment methods can already be applied successfully to these questionnaires as a potential source of structural validity. Also, design recommendations when constructing a FC questionnaire without a previous likert questionnaire and application are discussed considering empirical viability of simulation factors that seemed to improve dimensionality recovery, such as the inclusion of heteropolar or unidimensional blocks.

WEDNESDAY 3 JULY
Session 2.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

65. Preserving the dignity and the privacy of candidates in AI-based assessments

Nancy Tippins
The Nancy T. Tippins Group, LLC USA

Although AI-based assessments can often offer unobtrusive ways to evaluate candidates, such assessments also pose challenges to their dignity and privacy. The very use of AI-based assessment without human interaction may signal the lack of importance of the individual and affront a candidate's sense of self-worth. The input data used may further call into question the treatment of candidates. Job-irrelevant data in particular may suggest a selection system that is ineffective and decisions that are based on data that are neither fair nor useful. Privacy concerns may also diminish a candidate's self-esteem. Use of some data without informed consent can signal the employer's opinion of candidates in general, and storage of personal identifying information and other data may present not only concerns about the dignified treatment of candidates but also legal issues involving the collection, storage, and destruction of personal information. We will discuss ways in which a candidate's dignity is not preserved when AI-based assessments are used and the legal and professional requirements to protect the privacy of candidates and their information. We will conclude by presenting ways to mitigate some of the concerns about protecting the dignity of candidates and the privacy of their data.

WEDNESDAY 3 JULY
Session 2.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

89. From Job Descriptions to Facets: a Computational Technique for Effective Predictive Modeling

Tales Marra, Emeric Kubiak
AssessFirst

Objective. Personality importance in HR decisions is growing. However, it is essential to recognize that specific facets are advantageous in certain occupations and counterproductive in others (Judge ans Zapata, 2015). This study's goal is to devise an AI-driven method to identify relevant facets for an occupation. Method. First, GPT-3.5 Turbo was employed to analyze job postings and extract core skills. Second, a Sentence Transformer was used to generate relevant candidate facets corresponding to the evaluated competency. Third, we utilized DistilBERT, which was fine-tuned on the ESCO database comprising over 13,000 skills labeled with facets by a team of psychologists. This step discerns the direction in which a facet influences the competency. Fourth, to ensure gender neutrality, a neural network-based post-modeling correction technique was integrated. Results. We observed a 98% time reduction for predictive models creation. A task that required 45 minutes for human completion was accomplished by our model in merely one minute. Another study is ongoing, testing the predictive validity of the models. Conclusion. Our methodology diminishes recruiters' time investment, ensures job-specific facet relevance, and maintains gender neutrality, setting the stage for a more precise recruitment procedure.

WEDNESDAY 3 JULY
Session 2.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

264. The AI Will Interview You Now: Exploring Candidate Perceptions of Fairness in AI-Based Assessments

Karim Badr
SHL
Manuel Gonzalez
Montclair State University
Cam Beazley
SHL
Rudy Abi Habib, Pia Tohme
Lebanese American University

The utilization of artificial intelligence (AI) in assessment is rising. From automatic item generation to essay scoring, AI is permeating different aspects of test construction and administration. While AI can increase efficiency in personnel assessment and selection, concerns loom about its other implications. Among these concerns is how using AI may affect the candidate experience in asynchronous video interviews, such as the perceived face validity and fairness of AI-based scoring. Existing research findings on this topic are mixed. For example, research by Mirowska (2020) suggests that job candidates may be less likely to apply to an open position if the assessment process leverages AI, relative to an entirely human-led process. Researchers have also proposed and identified (albeit, with mixed evidence) potential factors that underlie candidates' reactions toward AI-based processes, such as reduced feelings of control (Gonzalez et al., 2022), and perceiving the process as depersonalizing (Gonzalez et al., 2019) or unjust (Acikgoz et al., 2020). Yet, data suggest that candidates may sometimes view AI-based assessment processes favorably, such as having enhanced accuracy and consistency (Nolan et al., 2016). In our research, we are investigating (1) how fair candidates perceive AI-based assessments to be, (2) whether these fairness perceptions differ based on candidate demographics (e.g.: gender, age, cultural background), and (3) what interventions can positively affect candidates' fairness perceptions. We are conducting an online experiment in which undergraduate students complete an AI-scored asynchronous video interview and report their perceptions of the process. We are manipulating whether (a) a pre-assessment explanation for using AI is provided and (b) whether participants are educated about AI via a post-assessment debriefing. Data collection for the study is ongoing at the time of this submission, and our full findings will be shared at the conference.

WEDNESDAY 3 JULY
Session 2.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

275. NMIRT: A neural network-based multidimensional item response theory model

Jiyuan Ding, Yan Cai, Dongbo Tu, Daxun Wang, Fen Luo, Zhichen Guo, Kai Liu, Qin Wang, Junhuan Wei, Fangbin Chen, Xuhong Song, Yuzhi Yan, Pan Jiang Jiangxi
Normal University

Given the multidimensionality of many tests, the widespread adoption of multidimensional item response theory (MIRT) has ensued. Nonetheless, the computation required for estimating MIRT models is substantial, contingent on the chosen estimation method, and there are typically constraints on the number of latent variables that can be estimated in the analysis. Marginal maximum likelihood estimation exhibits exponential growth in computational complexity during the marginalization process as the number of latent variables linearly increases. Similarly, the expectation-maximization (EM) algorithm also exhibits exponential growth in computational complexity with increasing latent variables. This study utilizes neural networks to apply the Rasch model in multidimensional item response theory, enabling the simultaneous estimation of a significant number of parameters simultaneously. Although these methods introduce more trainable parameters, these parameters can be seen as estimates of discrimination and difficulty parameters, among others. By optimizing a single loss function, all parameters are trained simultaneously. After training the neural network, all ability, discrimination, and difficulty parameters can be obtained. While this method employs a neural network model, it does not necessitate prior knowledge of the parameters to be estimated as training data. It only requires the response matrix of participants and the component matrix between dimensions and items for estimation. Simulation studies have demonstrated that this approach surpasses traditional algorithms, including EM, Metropolis-Hastings Robbins-Monro (MH-RM) algorithm, Monte Carlo EM, and quasi-Monte Carlo EM, in terms of estimation accuracy. Furthermore, it can handle the estimation of 15-dimensional test data, a task beyond the capabilities of traditional methods.

WEDNESDAY 3 JULY
Session 2.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

445. Assessment Vulnerability to Cheating Using Large Language Models

Richard Landers, Vivien Lee
University of Minnesota

Public, inexpensive large language models like ChatGPT and Bing AI are increasingly being used to cheat on cognitive assessments, especially in collegiate settings. In this study, all faculty in the United States' Minnesota State System (i.e., approximately 2000 faculty spread across 34 2-year and 4-year institutions and a variety of domains) were asked to provide a sample assessment from their course in any format (e.g., multiple choice, short answer, essay). Undergraduate research assistants having completed no prior courses in the discipline of the submitted assessment were then asked to use LLMs to generate their "best" set of answers to each assessment. Contributing faculty were then asked to grade that assessment as well as provide the mean and standard deviation of scores from the last time that assessment was used in their courses. In a subset of cases, faculty also provided a grading key, which enabled exploration of objectivity in grading when knowing AI was used. A subset of cases were also assessed by two research assistants to assess the interrater reliability of AI/LLM-based cheating. This project is currently in the midst of data collection, but the project is expected to be complete by April 2024.

WEDNESDAY 3 JULY
Session 2.7
Topic: Testing equivalence by psychometrics methods

54. Rethinking Psychometric Properties in Developing Instruments for Optimal Assessment

Joshua Chiroma Gandi
Department of Psychology

Psychometrics, being the science of psychological measurement, remains more increasingly relevant especially with commensurate due diligence to ensuring the sustained adequacies of its respective properties. Rethinking psychometric properties, particularly those commonly referred to as validity; reliability; and responsiveness, requires elucidating conceptual clarifications and evidential operational analyses fashioned in pursuit of most suitable coefficients with corresponding significance index for determining such properties. The available alternatives include Cronbach's alpha (α), Revelle's beta (β), Goodman-Kruskal's gamma (β), Guttman's lambda (λ), and McDonald's omega (ω) coefficients. This study focuses on suitable coefficient(s) for determining internal consistency significance levels which also have further implication(s) in measuring validity and the scale items' responsiveness. A meta-analysis of equivalence tests with reference to significance levels corroborates that α focuses on internal consistency, β focuses on relationships between predictor and outcome variables, λ is a correlation factor which measures association for ordinal variables, λ measures the strength of relationship between two nominal variables, and ω emphatically determines internal consistency reliability. It clearly shows that α and ω are the principal indicators of internal consistency. The study found that α is easily bedeviled by confounding variables for being based on correlation, built on the assumption that responses are normally distributed and equally explains the factor. Despite associated disadvantages including type I error (false negative) which rejects what should be accepted just like β with type II error (false positive) which accepts what is reject-able, α is more frequently being used than ω due to its simplicity and accessibility. This study devised a methodically easy, friendly, most effective and adaptable way of computing ω for ensuring psychometric properties.

WEDNESDAY 3 JULY
Session 2.7
Topic: Testing equivalence by psychometrics methods

83. Investigating re-sitting effect on item difficulty in a medical selection test

Luc Le
Australian Council for Educational Research

Introduction Health Professions Admission Test – Ireland (HPAT), developed by the Australian Council for Educational Research (ACER), has been used in the selection process for applicants to Medicine courses in some universities in Ireland. HPAT consists of three separately multiple-choice section: Logical Reasoning and Problem Solving; Interpersonal Understanding; and Non-Verbal Reasoning. The sections were designed to measure cognitive abilities and skills that have been found useful for identifying candidates who might be suitable for medical study. Some candidates can participate in two consecutive yearly testings. Objectives This study was designed to investigate the effect of the re-siting group on the item difficulty estimates and on the inter correlations among the test section scores. Design/Methodology HPAT 2023 Data was used, included 3105 candidates with 615 (20%) of them already did HPAT 2022. The Rasch model was implemented to analyse the response data for each test-taker group in each HPAT section. Differential item functioning (DIF) was computed as the difference between the item difficulty estimated from the re-sitting group and that from the other test takers. Additionally, inter correlations among the test sections for the re-sitting group and other candidates were computed and compared using Cohen's effect sizes (Cohen, 1988). Results & Conclusions Main results showed a minimal or small Cohen's effect sizes in the difference between the inter section correlation from the re-siting candidates and from other candidates. Moreover, there was only one or two items with large DIF in each section. In conclusion, there would be a very slight impact of performance of re-sitting candidates on the item difficulty and inter correlations among the HPAT section scores. References: Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillside, NJ: Lawrence Erlbaum Associates.

WEDNESDAY 3 JULY
Session 2.7
Topic: Testing equivalence by psychometrics methods

523. Fostering Fairness: Assessing measurement equivalence of MMIs

Mustafa Asil, Amy Bannatyne, Belinda Craig, Kirsty Forrest, Jessica Stokes-Parish, Jaclyn Szkwara
Faculty of Health Sciences & Medicine, Bond University, QLD, Australia

Introduction The selection process for medical schools plays a critical role in identifying candidates who possess the necessary attributes and competencies to excel in the demanding field of medicine. This Multiple Mini Interviews (MMIs) have gained popularity as an assessment tool for evaluating applicants' non-cognitive skills, such as communication, empathy, and ethical judgment. Ensuring the fairness and validity of MMIs across diverse applicant groups is paramount for equitable selection processes. This study aims to investigate the extent to which: · core attributes assessed by MMIs maintain consistent meaning and interpretation across gender groups. · gender-related disparities do exist in MMI performance. · examiner gender influences student MMI performance. Method This research analyses data collected from 2024 Bond university medical school applicants. A multigroup confirmatory factor analysis (MGCFA) approach will be employed for a comprehensive evaluation of the underlying construct validity of the MMI, assessing its consistency across different subgroups. Results This study holds promise for shedding light on the extent to which MMIs are comparable and unbiased across different genders. By comparing the performance of female and male students, the analysis provides insights into potential variations in non-cognitive skill assessment between genders. Additionally, the analysis will investigate whether the gender of examiners influences student MMI outcomes, addressing the potential for examiner-related biases. Discussion/Conclusion By unveiling potential measurement biases and inequities, this research contributes to the refinement of admissions procedures, ultimately fostering a more inclusive and just selection process for aspiring medical professionals. These findings hold implications for both medical education policymakers and practitioners seeking to ensure a more inclusive and equitable selection process for aspiring medical professionals.

WEDNESDAY 3 JULY
Session 2.7
Topic: Testing equivalence by psychometrics methods

611. Studying invariance with multiple groups in large datasets: A comparison of bayesian SEM and alignment method

Oscar Lecuona
Universidad Complutense de Madrid
Ricardo Rey-Sanchez
Universidad Autónoma de Madrid
Manuel Martín-Fernández
Universitat de València

Conceptual framework: Cross-cultural research often deals with estimation of latentodels across multiple groups, and ideally, large datasets. This can introduce bias due to cultural variations in the item parameter estimates of instruments. Classic invariance models often are too strict to study if these properties are stable across groups due to their tendency to reject the invariance hypothesis when the number of groups increases. A potential solution to this issue is to consider approximate measurement invariance, which identify sources of non-relevant variations across groups and allows some wiggle-room in the item parameter estimates of each country. Techniques that feature this framework include bayesian structural equation modelling, and the alignment optimization method. The first one allows for variations between groups to be modelled with expected variations (e.g., zero-centered normal distributions). The second estimates the structural parameters that minimize the differences between the measured parameters across groups through an optimization algorithm. Objectives: We apply BSEM and AM to a dataset of multiple groups and large number of observations to (1) illustrate each technique with an applied example, and (2) compare both techniques to dillucidate strengths and limitations. Sample: A multi-country sample of 7,254 participants and 10 countries: All sub-samples but one were predominantly middle-aged (around 40 years), while also female (around 60%) except two sub-samples. All sub-samples measured optimism with the Life Orientation Test-Revised, with 10 items and a two-factor latent structure of optimism and pessimis, and an alternative structure of one overall optimism factor. Implications: Both methods provide pathways to study invariance in psychometric instruments when dealing with multiple groups. However, each one provides a series of advantages and caveats that need to be considered. We discuss initial recommendations for these scenarios

WEDNESDAY 3 JULY
Session 2.8
Topic: Quantitative, qualitative, and mixed validation methods

177. A Conceptual Framework for Assessment Experience in Educational Testing.

Fernando Mena Serrano
University of Massachusetts Amherst
Sergio Araneda
Caveon Test Security
Stephen Sireci, Eduardo Crespo Cruz
University of Massachusetts Amherst

In this paper, we present an framework to characterize assessment experiences in educational testing. However a number of authors have done research in the last ten to twenty years in topics related to user experience in educational testing, like test-taker engagement (Wise, 2017; Wise & Kong, 2005; Wise & Kuhfeld, 2021), socio-cognitive aspects (Mislevy, 2018), the inclusion of test-taker voices (Cheng & DeLuca, 2011; Gardiner & Howlett, 2016; Song, 2018; Xie, 2011), or through the concept of test-taking experience (Burstein et al., 2021); in none of the above there is a definition of the concept of experience. Our proposed framework is built upon John Dewey's three main principles of experience (Dewey, 1938), alongside methodologies from user experience (UX) research (Forlizzi & Battarbee, 2004; Holbrook & Hirshman, 1982), and adapts these to the educational testing landscape. We delineate five levels of sense-making: Free-stream, Immediate, Integrative, Co-integrative, and Dialogical; five types of interactions between an examinee and a test: Fluent, Cognitive, Emotional, Expressive, and Physical; and six stages relevant to the assessment experience: Ante, Pre, Zero, Inter-post, Post, and Long-term experience. We analyzed tweets shared by examinees about a university admission test in Chile. Two raters identified interactions on tweets written by examinees. We calculated Phi values for the binary variables indicating types of interactions in the tweets, as well as Cohen's Kappas. We find Kappa values above 0.6 for cognitive interactions, values between 0.45 and 0.33 for emotional, expressive and physical interactions; and null inter-rater reliability for fluent interactions. In conclusion, we argue that our conceptual framework offers a comprehensive understanding of user experiences in testing, enhancing test design and validation processes. We advocate for further research to extend this work and improve educational and psychological testing practices.

WEDNESDAY 3 JULY
Session 2.8
Topic: Quantitative, qualitative, and mixed validation methods

201. **The effect of starting with easy items on SEM in a CAT**

Serkan Arikan
Bogazici University
Eren Can Aybek
Pamukkale University
Gunes Ertas Polat
Bogazici University

As CAT is a comparably new testing system, the research on improving its functionality is growing. One of the issues in CAT is selecting the first items. Starting with collateral background information about examinees is a way to initialize the CAT and this approach could give better estimation of ability level of examinees. This study aimed to develop a CAT to assess 4th grade students' mathematical abilities and test the effect of starting with a very easy items on standard error of measurement (SEM). It was hypothesized that if anxious students start with easier items, their ability level would be estimated with higher precision in CAT. Therefore, the research question of the study was "keeping test length fixed in a CAT application, what is the effect of starting with a set of easy items on SEM for state test anxiety groups?" Overall, a 540-item pool of mathematics items was developed. Then, these items were calibrated with 1-PL IRT by administering to 3108 students as a pilot administration. After evaluating the item parameters, the live CAT administration system was constructed on Concerto platform. Additionally, state test anxiety scale was developed to be able to measure student anxiety level just before a test. Initially, 12 statements regarding the state test anxiety were developed and then piloted and the final form with 9 statements was created. For the real CAT administration, 403 students took a state anxiety test and then got a live CAT. Students were randomly assigned to 2 different groups. Two-way ANOVA results indicated that, controlling test length, there was no interaction between anxiety levels and starting a test with easy items on their SEM. Also, comparing anxiety groups, no difference on SEM was found. However, there was a significant difference between SEM of students. The effect size was found to be large. This finding shows that when students start with easy items, less SEM was observed regardless of their anxiety level.

WEDNESDAY 3 JULY
Session 2.8
Topic: Quantitative, qualitative, and mixed validation methods

549. Utilizing experts' judgments to determine cut-off scores of a test with little data

Julien Mouchnino, Dominique Casanova
Le français des affaires (CCI Paris Ile-de-France)

Adapting evaluation instruments to minority languages and cultures leads to a fragmentation of the target audience. In such conditions, collecting large enough pre-test sample data may be challenging. An alternative is to collect expert judgments on items difficulty, such as what can be implemented within the framework of some standard setting procedures (e.g. Angoff). But the quality of direct prediction of item difficulty by experts is sometimes questioned (Hambleton et al., 2003). In this paper, we study the extent to which expert judgements can be combined with limited pre-test data to improve the prediction of items difficulty. We start with the judgements of 16 experts collected during an experiment designed to align the Test d'évaluation de français with the Common European Framework of Reference. We compare mean experts judgments with the empirical difficulty of the items (all dichotomous). Based on simulations, we estimate the sample size required to obtain a similar prediction quality by pre-test, as well as the improvement in prediction that can be expected by combining expert judgements and pre-tests on small samples. The results show that expert judgements cannot replace pre-tests on 200 individuals to predict item difficulty. They are nevertheless of undeniable interest when the capacity to gather empirical information is limited. In particular, the combination of expert judgements and estimates obtained from small samples helps to improve the quality of the prediction. The use of anchor items of known difficulty can help identifying a relevant ratio for the combination of the two types of predictions (expert judgments and pre-test estimates). However, a posteriori analyses are necessary to assess the quality of the procedure used.

WEDNESDAY 3 JULY
Session 2.8
Topic: Quantitative, qualitative, and mixed validation methods

711. Recent Trends, Emerging Controversies, and Consequences for Response Processes Validation: Lessons Learned

Anita M. Hubley, Sophie Ma Zhu
University of British Columbia

Framework: Response processes (RP) include what people think, feel, and do when responding to items. Examining the fit between actual and theoretically expected processes addresses a key source of validity evidence in the Standards for Educational and Psychological Testing (AERA et al., 2014). Traditionally, there have been few RP validation studies in the literature. Objectives: The purpose of this presentation is to describe some recent trends in RP validation, some emerging controversies, and potential consequences for the future of RP as a source of validity evidence based on our experiences conducting and examining such research. Results: Recent trends include a notable increase in RP validation work alongside increasingly diverse reasons for exploring RP. As a result, there are emerging concerns and controversies. A lack of detail and clarity in the reporting of procedures often makes studies difficult or impossible to evaluate or replicate. Critically, researchers rarely identify, a priori, the intended RP to which they could compare actual RP and, thus, neglect to establish an argument for the validity of score inferences. Researchers sometimes simply equate the methodology used (e.g., cognitive interviews or CIs) with RP validation. Unintended consequences of using common methods such as CIs to identify problems in item comprehension and wording may be a conflation of test development and test validation, and confusion between RP validation vs. test content validation. Consequences like these can erode the value of RP as a source of validity evidence. Implications: RP evidence has enormous potential for informing the validity of inferences made from item and test scores. It is critical to identify and address misunderstandings, poor practices in reporting and evaluating RP evidence, and other concerns if this source of validity evidence is to reach its full potential and not devolve into an activity that lacks meaning and value.

WEDNESDAY 3 JULY
Session 2.8
Topic: Quantitative, qualitative, and mixed validation methods

798. Examining undergraduate students' programming process through cognitive interviews and keystrokes

Min Li
University of Washington

Learning to code is becoming not only a popular subject for students and professionals but also a critical literacy in modern society. Yet, research is nearly unanimous that computer programming is difficult to teach and assess (Fincher & Robins, 2019). This small-scale study of examinees' keystroke data and interviews addresses such a challenge by exploring how student programming strategies and processes are associated with task complexity and prompt scaffolding. Extending research on keystroke log analysis of natural language writing and computing education, we ask two research questions: (1) how do task complexity and scaffolding of prompt affect student programming process? (2) what types of struggles and programming strategies are observed with students varying in programming proficiency and demographics (i.e., ethnicity, gender, and home language)? A total of 48 undergraduates from introductory Python undergraduate courses in the US and Canada were included in this study. The keystroke data on 21 coding tasks was collected via a platform capable of capturing mouse and keyboard actions (and timing). After completion of programming tasks, each individual was interviewed in a one-on-one Zoom, based on a retrospective interview protocol, asking participants to explicate their cognitive and metacognitive processes by clarifying codes, describing struggles and/or mistakes, and explaining programming strategies. Qualitative analysis of interviews and quantitative analysis of keystrokes revealed interesting relations between task characteristics and process features, such as pauses between lines, and time spent on editing and revising codes. This study is expected to address validity and fairness issues around computing assessment: (1) little is known about what keystroke features can be used to differentiate student performance varying in programming proficiency; and (2) it is unclear what item characteristics may influence individuals' programming processes.

WEDNESDAY 3 JULY
2.9 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

233. The assessment of cognitive abilities: critical approaches to validity and reliability (233)

Vanessa Torres van Grinsven
Open Universiteit AND University of Cologne

Standardized tests, such as intelligence tests, are commonly used diagnostic tools. They play an important role in clinical decision making and allocation of education and healthcare. Research has however shown that important errors may occur despite the application of validation processes and adherence to quality criteria for psychometric tests. In this symposium we will present a few critical approaches to validity and reliability in standardized testing. The interpretation of a test score relies on information gained from group data and, often, tests developed based on classical test theory. It will be argued that group data can´t provide justifiable information about the individual, and this may have consequences for the wellbeing of the individual. A shift in perspective arises when individuals are followed over time. Thus, differences in between-person data and within-person data will be discussed. We delve into the theoretical foundations of intelligence tests and the implications these have for validity and discuss objections from the professional field regarding the dominant role of IQ test scores in Dutch policy, based on a recent questionnaire among practitioners. From another viewpoint, it will be discussed that test scores may be biased in relation with unsolved issues of multidimensionality that have to do with, among others, personality diversity and differential skills. From an interactionist and process-performance approach, pre-test methods, qualitative methods stemming from the field of survey methodology, will be proposed as a way to approach the problem of multidimensionality in tests. Finally, Dynamic Assessment will be presented as an alternative to assess children, including a discussion of advantages and disadvantages and the challenges and opportunities DA can offer for health and school services. DA has shown to provide insights into the most appropriate instructional level and increases insights in capabilities of children and adults.

Discussant name:
Discussant surname:
Discussant affiliation:

WEDNESDAY 3 JULY
2.9 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

234. The Response Process in Standardized Cognitive Ability Tests and Validity

Vanessa Torres van Grinsven
Open Universiteit AND University of Cologne

Standardized cognitive ability tests, such as the Wechsler Intelligence Scale for Children (WISC), are common diagnostic tools used in the process of diagnosing learning and behavior disabilities. Important decisions are made based on the results of tests scores, with little consideration for, for instance, emotional variables that might impact negatively a child´s ability to perform up to his or her potential on standardized tests. In short, standardized tests may be biased in relation to social background and cultural and ethnic diversity, but also personality traits and motivational and emotional factors (personality diversity) and differential skills. In my presentation, I will give a review of empirical research that places attention on two types of bias in intelligence testing with the WISC: bias occurring related to diversity in personality traits and emotional and motivational factors that interact with test characteristics, and bias related to the influence of the assessor. I will then discuss that psychometric procedures seem not to be able to solve this problem. This could be related to the unsolved issue of multidimensionality. Multidimensionality can be approached from the viewpoint of the response process: in a test-taking situation with a standardized test, complex interrelationships take place between the test-taker and the test, and an assessor if present. This complex interaction shapes the response process which results in a performance, i.e., the results of this test. I present the interactionist and process-performance approach as a framework to approach this issue, and propose the “pretest methods,” qualitative methods stemming from the field of survey methodology as a way to approach the problem of multidimensionality stemming from the response process and help improve the development and interpretation of results of standardized tests.

WEDNESDAY 3 JULY
2.9 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

235. **Intelligence and the Individual**

Anouk van Hoogdalem
Radboud University

Based on the policy in the Netherlands, it appears that IQ plays a significant role in clinical decision-making. IQ functions as a barrier, a phenomenon often referred to as 'doorway diagnostics'. The scores determine to a considerable extent the allocation of education and healthcare. In other words, IQ-scores seem to be 'magical numbers'. On the other hand, objections to this approach are raised from both the scientific community and the professional field. We delve into the scientific objections against the use of IQ as a doorway statistic, examining the validity and reliability of standardized intelligence tests and their value for the individual. The prevailing understanding of intelligence, including its conceptualization and measurement, largely relies on group-based research. A shift in perspective arises when individuals are followed over time. I will therefore present the differences between between-person data and within-person data. Besides, the (assumed) theoretical foundation of intelligence tests, like the CHC model, and the implications this has for validity will be outlined. Additionally, I will address objections from the professional field regarding the dominant role of IQ-scores in Dutch policy, based on a recent questionnaire among Dutch pedagogues and psychologists. The foundations of these objections and the desires for an alternative diagnostic approach will be explored.

WEDNESDAY 3 JULY
2.9 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

244. **The Test Tested**

Anna M.T. Bosman
Radboud University

A scientist-practitioner, who interprets a test score on a psychological test to make a reliable statement about an individual, has to rely on information gained from group data. Test developers take into consideration measurement errors that rely on probability and error theory originally developed in astronomy and physics. The average of the distribution is interpreted as the most reliable measurement value. I will argue that this procedure is fundamentally flawed. Group data cannot provide justifiable information of the individual because human processes do not obey the Ergodic Theorem. Ergodicity refers to two principal requirements: 1) homogeneity, which refers to the? fact that each subject in the population must obey the same statistical model; for example, a factor model with its loadings must be invariant among individuals, 2) stationarity, statistical parameters should remain invariant irrespective of time and place; in other words, the statistical properties of developmental processes should not change over time. That, however, violates the assumption of development. My goal in this presentation is twofold, I will explain 1) how classical test theory led to the development of psychological tests that cannot withstand the test of criticism 2) the consequences of this practice for the individual.

WEDNESDAY 3 JULY
2.9 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

394. **Dynamic assessment; An alternative to static testing?**

Maartje Radstaake
Radboud University

Children with developmental disabilities are often different on many aspects, for instance medical background, functional abilities and levels of developmental stimulation. These differences create obstacles for development. When these obstacles can be removed, abilities can be better assessed. In dynamic assessment, exactly this is done. Following the guiding principles of Feuerstein and Vygotsky, dynamic assessment assesses how a child learns by closely examining a) which cognitive functions lie within and outside the Zone of Proximal Development and b) how cognitive functioning can be strengthened through mediation. Dynamic assessment gives insights in the most appropriate instructional level and this approach has shown to increase (insights in) capabilities of children and adults with o.a. autism, developmental disabilities, differing cultural background and high-functioning children. Criticism on dynamic assessment mainly revolves about the time consuming nature of the assessment, the needed skills and capabilities of the assessor and challenges with validity and reliability of the assessment. But how to assess reliability and validity when development is unstable and dynamic in nature and highly dependent on context, inter and intrapersonal factors, as we learned in the previous workshops? Do we wish to predict development or to learn how to defy this prediction? During this workshop, we will held a short discussion about the challenges and opportunities DA can offer for health and school services.

WEDNESDAY 3 JULY
Session 3.1 SYMPOSIUM
Topic: Innovations in test development

164. Implementing a New Model for Measuring Clinical Judgment in Nursing: The Next Generation NCLEX

Joe Betts, William Muntean
NCSBN
April Zenisky
University of Massachusetts Amherst
Shu-chuan Kao, Doyoung Kim
NCSBN

This symposium will highlight the research of a large-scale testing program's efforts to expand the exam to include the complex construct of clinical judgment (CJ). The Session will describe the evolution of this research project from inception to implementation. The first paper will discuss the guiding R&D model used for the project, Figure 1. Then the foundational research that led to the development of the cognitive conception of clinical judgment measurement model will be discussed (Muntean, 2012). This will highlight foundational research spanning from a nationally representative strategic job task analysis to the use of a linkage methodology to cross classify entry-level nursing knowledge, skills, and abilities to the final development of the item and scenario writing task model (Betts, et al. 2019). The second paper will focus on research related to item scoring research and validation studies that are appropriate for analyzing evolving case studies (Betts, et al., 2022). The third paper will focus on work leading up to the final implementation of the polytomous computerized adaptive test (CAT) development (Kang, et al, 2022a; Kang, et al., 2022b) and will highlight the research and testing that was done to prepare for the initial beta-test and culminated in the final public deployment of the exam. Information and methods discussed can generalize to any program investigating extending their exams or measuring complex constructs. The final paper will provide a perspective from a Technical Advisory Board Member who oversaw all aspects of the project from inception to implementation.

Discussant name:
Discussant surname:
Discussant affiliation:

WEDNESDAY 3 JULY
Session 3.1 SYMPOSIUM
Topic: Innovations in test development

176. The Best Seats in the House - A Viewpoint on the Development of the Next Generation Nursing Examinations (Validity theory in testing, psychological assessment and survey research)

April Zenisky
University of Massachusetts Amherst

Imagine for a moment you are at a Taylor Swift concert or a Real Madrid football game, but your seat has been upgraded – rather than sitting in the grandstands, you've been asked to sit in a premium suite. You're joined by several other superfans, as well as a few pros who have been deeply involved in the show unfolding in front of you. During the show, you learn all the fascinating behind-the-scenes details and the pros even ask for your input periodically. You're also treated to a remarkable meal, and a delightful intellectual camaraderie develops among everyone present. That's what it's like to be on the NCLEX technical advisory committee (TAC). Serving on the TAC for the NCLEX really is akin to fabulous seats at an incredible performance venue, but one where you are also invited to participate in the production. My perspective is not only overseeing the crafting of evaluation metrics but also witnessing the evolving nature of the nursing profession. Our committee plays an key role in ensuring that future nurses are rigorously tested for their clinical acumen, theoretical knowledge, and empathy-driven patient care. It's not all fun and games: TAC members have a responsibility to help ensure the relevance and quality of the exam given rapid medical advancements and psychometric development, while preserving age-old virtues of nursing. In this presentation, I will trace some of the highlights of the development of the Next Generation Nurses (NGN) Examination over the past 6-7 years, from the perspective of a member of the TAC. I will reflect on various aspects of the test development process, such as item development, calibration, timing analyses, standard setting, and the evolution of the validity agenda, with the focus being on the interplay of research done by psychometric researchers at the National Council of State Boards of Nursing (NCSBN) and the conversations that ensue during TAC meetings, within and between members of the TAC and NCSBN staff.

WEDNESDAY 3 JULY
Session 3.1 SYMPOSIUM
Topic: Innovations in test development

166. Developing and Testing Scoring and Response Models for Clinical Judgment Case Studies (Psychometric modeling)

Joe Betts, William Muntean, Shu-chuan Kao, Doyoung Kim
NCSBN

This presentation will discuss the research used to validate the Scoring and Mathematical Model aspects of the comprehensive R&D model to explore measuring clinical judgment (CJ) in nursing, see Figure 1. Numerous research studies will be described and results provided that substantiated a set of scoring models for the new polytomously scored, technology enhanced item (TEI). The results of these experiments lead to a set of raw scoring methods that effectively measured the items for partial credit (Betts, et al., 2022). Research on applying different item response models (IRT)will be discussed. This will bridge the gap between computing the raw score on the item and the mathematical function used to scale the items to the underlying score scale. IRT and confirmatory factor analytic will be discussed that validate the underlying factor structure. An interesting aspect of the research will also be the unique methods employed to evaluate evolving case scenarios which are a set of related items each measuring a specific cognitive element of the evolving case scenario focused on eliciting good decision-making skills, see Figure 2. Figure 3 provides an overview of one of the unique methods for treating item sets as a super-polytomous model. The importance of this research was to help evaluate the extent to which item sets have between item dependencies and the need for a testlet structure for measurement (Kang, et al., 2022a; Kang, et al., 2022b). Discussion of these results will be provided.

WEDNESDAY 3 JULY
Session 3.1 SYMPOSIUM
Topic: Innovations in test development

167. Implementing a Polytomous CAT with Evolving Case Studies (Psychometric modeling)

Joe Betts, William Muntean
NCSBN

This paper will focus on organizing all the information from the comprehensive R&D model (Figure 1) into the final Assessment model used in practice to measure clinical judgment. This resulted in the first large-scale licensure examination employing a polytomously scored computerized adaptive testing (CAT) methodology. Along with the implementation of the polytomous scoring, this paper will also discuss how it is the first exam to include item sets, called evolving case scenarios, into an adaptive exam. The previous papers in this symposium set the foundation for the tools available to the CAT delivery, and this paper will discuss how those tools and results were put together to implement a polytomously scored CAT that incorporates the use of item sets. There were a number of simulation studies employed to help guide the development of the final design that will be discussed (see for example, Kang, et al, 2022a and 2022b). In addition, the numerous operational challenges and tasks needed to transform a licensure exam from its legacy format to the new polytomously scored CAT will be discussed. Methods and procedures for testing and validating prior to the operational release will be discussed. A discussion of the end-to-end psychometric testing scenarios will also be discussed which were used to validate test driver scoring and examine ridge-case scoring situations. All the development work and testing resulted in an error-less operational launch and understanding these principles can help any test development team in a similar circumstance.

WEDNESDAY 3 JULY
Session 3.1 SYMPOSIUM
Topic: Innovations in test development

198. Development of a Conceptual, Task, and Assessment Model for Measuring Clinical Judgment (Psychometric modeling)

William Muntean, Joe Betts, Doyoung Kim, Shu-chuan Kao
NCSBN

Figure 1 provides an outline of the R&D model used to validate the measurement of clinical judgment (CJ) skills in entry-level nursing. This presentation will provide information related to the Conceptual and Task Model components. The foundational literature review of the importance of CJ for nursing (Muntean, 2012) indicated that deficits in CJ were related to a number of negative patient outcomes suggesting the need to ensure entry-level minimal competence was important. The review also provided a number of information processing and problem-solving strategies utilized in nursing education. Taken as a whole, this information led to the development of the clinical judgment measurement model, Figure 2. To future validate the necessity of CJ for entry-level practice, a strategic job task analysis (SJTA) was undertaken. The SJTA was comprehensive as it involved a number of different research projects. For instance, one research project utilized a linkage methodology (Raymond & Neustel, 2006) that allowed for the cross classification of knowledge, skills, and abilities being evaluated. Figure 3 provides an overview of the highest knowledge statements and skills that underly the 286 entry-level task statements. Other research projects and all results will be discussed. This presentation will also provide background information that bridges the gap between the validation of the CJ construct and the scenario development process needed to measure CJ. In Figure 1, this is represented by the Task Model used to finalize the item development process for developing evolving case studies (Betts, et al., 2019). Numerous research endeavors will be discussed outlining how this was accomplished. The final process for building and reviewing case studies exemplifies how the Conceptual and Task models can be combined into a comprehensive development methodology. This method is not just restricted to nursing but is highly generalizable for any program.

WEDNESDAY 3 JULY
Session 3.2 SYMPOSIUM
Topic: International assessment

511. Insights on changes in test use across Europe from the Work of the EFPA Board of Assessment

Nigel Evans
NEC

The Board of Assessment (BoA) has the direct reach to impact test use practice for almost half the practicing Psychologists in the world. It is one of the oldest and largest working groups of the European Federation of Psychological Associations (EFPA). This symposium will share background, overviews, and insights on the changing nature of test use through various BoA member initiatives. The aim of the BoA is to develop and improve assessment practice in Europe, and beyond. To this end, the BoA have published a review process known as the Test Review Model (TRM) which is extensively used as a standard across Europe in evaluating tests. Similarly for test user standards, there is the Euro Test certification process for meeting expected competence in administration, feedback, and integration of tests within common assessment domains. Surveys are also conducted on attitudes to testing and test use. To understand changing practices in assessment, BoA members will share the impact of their work across four different countries, essentially to improve quality in assessment and effectively lobby for better testing practice. European Psychologists involved in assessment are encouraged to engage with the work of the BoA by directly accessing general guidance and specific standards documentation through their member representative and EFPA website.

Discussant name
Discussant surname
Discussant affiliation

WEDNESDAY 3 JULY
Session 3.2 SYMPOSIUM
Topic: International assessment

566. Online administration of tests of Italian psychologists: attitude and behaviours at the time of COVID-19 within an European Federation of Psychologists' Survey (Validity theory in testing, psychological assessment and survey research)

Adriana Lis
Università di Padova
Daniela Traficante
Università Cattolica del Sacro Cuore
Andrea Bobbio
Università di Padova
Filippo Aschieri
Università Cattolica del Sacro Cuore

Notwithstanding the debate about the advantages and disadvantages of internet administration of tests and testings, little is specifically known about recent attitudes concerning this issue in Italy, above all around March 2020 when Italy was under rigorous limitations on socioeconomic activities and movements (Ministero della Salute, 2020) as a response to the COVID-19 pandemic. This presentation aims to add information to this topic. 828 Italian psychologists all using tests participated in the study extracted from a larger sample. Data for this study derived by (a) Recent data derived from a European Federation of Psychologists' survey about agreement about the advantages of internet administration, improvement of administration quality, the possibility of fraud, privacy, and quality of test administration, (b) A 12-item questionnaire concerning various issues about online test use (e,g, similarities and differences between online administration, scoring, normative data; adaptation according to different kinds of clients, etc.); (c) specific information about how psychologists managed the use of testing in this specific situation (phone, zoom etc.) and how clients reacted.

WEDNESDAY 3 JULY
Session 3.2 SYMPOSIUM
Topic: International assessment

614. Revitalizing EFPA EuroTest Standards: Possible Strategies for Integrating Test User Standards into the EFPA Information Strategy (International assessment)

Urszula Brzezinska
Pracownia Testow Psychologicznych PTP
Nigel Evans
Director NEC, UK

The EFPA Board of Assessment developed the EuroTest standards for qualifications in test use to foster an advanced testing culture within EFPA member states. Alongside the EuroPsy and the Test Review Model, the EuroTest standards constitute a key component of the EFPA information strategy within the test use context. This strategy aims to enhance the psychological testing market by ensuring top-tier quality facilitated by competent professionals. The key aim of the EuroTest standards is to circulate a European-level benchmark, providing a reference for comparing local national systems. Upon recognition by EFPA, the local certification system for test use attains an equivalent status. The EFPA test use model strives to establish European standardization for the criteria a test user needs to meet when acquiring a psychological test from its publisher, making it a highly welcomed case. EuroTest enables the implementation of a standardized assessment for test users' competences, supporting national psychological association in shaping their polices. This includes establishing European standards for education, professional training, and competence in psychology. Moreover, EuroTest plays a crucial role in setting security protocols for accessing psychological tests and shaping policies for their safe and responsible use in various socially sensitive applications of psychological diagnosis. The presentation will employ a three-showcase analysis to emphasize the significance of introducing EuroTest as a bridging element for the EFPA information strategy. By juxtaposing EuroTest with EuroPsy Advanced Specialization, ISO 10667, and the Test Review Model, it will underscore EuroTest's crucial role in recognizing professional diagnostic competence as a core aspect of the psychology profession. Consequently, the audience will be encouraged to contribute to EFPA BoA initiatives aimed at promoting a European benchmark for a transparent testing culture.

WEDNESDAY 3 JULY
Session 3.2 SYMPOSIUM
Topic: International assessment

649. How can we contribute to improving the quality of psychometric practices in the field of guidance and work: attempts, actions undertaken and reflection underway in France (Validity theory in testing, psychological assessment and survey research)

Even Loarer
Cnam - Inetop, France

In France, as in other countries, the use of tests is very frequent, whether for self-knowledge and help in choosing a career path or for recruitment and detection of potential in companies. The development of digital technology in this area has led to a proliferation of operators offering online evaluation platforms and services. However, today there is no control of the quality of the services and tools offered. We looked into this subject to note that few publishers provide validation elements for the tests they distribute or the evaluation services they offer and the quality of these tests and services is very uneven, with some being of really bad quality. We have identified several levels of analysis of the quality of these practices: 1. Quality of practices of test publishers (ethics, commercial policy, sensitivity to ITC guidelines, collaboration with research laboratories, etc.) 2. Quality of tests (sensitivity, reliability, different forms of validity, norms, existence of external evaluations, etc.) 3. Quality of test use (suitability for the purpose and target population, conformity and standardization of implementation, rating, interpretation and restitution of results, qualification of users, etc.). These three aspects constitute what we call the "validity chain of psychometric practices" because the three components must be strong for a quality evaluation. Actions have been tested or are being implemented which aim to strengthen these three aspects: - Creation of a charter for test publishers proposed for signature by publishers, - Protocols for analysis of test quality, - Training of test users and awareness of buyers of tests and evaluation solutions of the issues and quality criteria in this area. The objective of the communication is to present these actions and the pitfalls encountered, and to share the experiences of other countries on this subject.

WEDNESDAY 3 JULY
Session 3.2 SYMPOSIUM
Topic: International assessment

771. Challenges and Insights to Psychological Testing within Forensic Contexts in the UK (Validity and fairness in cross-cultural testing, psychological assessment and survey research)

Glenda Liell1/1
CTS, British Psychological Society
Nigel Evans2/2
Director NEC, UK

Psychological tests and assessments are used in a variety of forensic contexts where the client is likely both vulnerable and powerless. Whilst UK Standards of Test Use have been applied over many years (through BPS EFPA Level 1&2 qualifications), practices in the use of tests have changed with the emerging evidence, as well as the broader use of alternative approaches such as a Structured Professional Judgement (SPJ's) tools and formulation-lead assessments. Some of the criticisms of psychometric measures which arguably most troubles practitioners, are that which concerns their applicability to an extremely heterogenous client base. In recent years there has been a growing interest in exploring the extent to which biases exist in the testing process. This extends not only to the assessment tools being used, but to the practitioner themselves, as well as broader concerns about biases in criminal justice systems. Indeed, the very notion of risk and risk factors and their conceptualisation has been called into question. This has led to a drive towards a more individualised and culturally informed approach to assessment, which can identify, in a flexible way, a multitude of potential issues contributing to offending behaviour. The presentation will briefly explore some of the above issues and pose some questions around the potential for the use of psychometric tests in forensic contexts to decrease, in whether ongoing development of tests is keeping pace with the demands for culturally informed assessments, what the contribution of psychometric testing can be in specifically individualised assessment, and to what extent are test developers and publishers attuned to shifts in the evidence and practitioner-thinking. Insights will be shared to inform similar testing shifts in practice across other areas of applied psychology.

WEDNESDAY 3 JULY
Session 3.3 SYMPOSIUM/PANEL
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

56. Organized Discussion: Rethinking Conceptions of Fairness and Equity in Assessment (56)

Randy Bennett
ETS

Many countries are experiencing population shifts resulting in unprecedented levels of diversity. With respect to race/ethnicity, the US has changed from 80% Caucasian in 1980 to 59% in 2021, a doubling in the percent of people of color from ~20% to ~40%. As of 2018, US public schools were 53% students of color. Countries throughout Europe, as well as on other continents, are experiencing similar shifts. These shifts encompass many human characteristics which inevitably bring greater levels of cultural diversity. For testing in education and in the professions, cultural diversity poses a challenge because tests are cultural artifacts that privilege certain ways of knowing, modes of expression, forms of representation, and linguistic structure. The standardized tests we use today have not evolved substantively to the same degree as populations have changed. Recognizing these needs, multiple approaches have been proposed, including assessments that are culturally responsive, justice-oriented, socioculturally responsive, and antiracist These approaches differ but share at least one key proposition: Designing assessment for the social, cultural, and other relevant characteristics of individuals and the contexts from which they come should enhance equity. The goal of this Session is to facilitate an organized discussion among four panelists (Maria A. Ruiz-Primo, Jennifer Randall, Ye Tong, Randy Bennett) and with the audience. Using the US and other countries as examples, the panelists will address such questions as: · Why does population diversity matter for testing and assessment? · How may traditional approaches fail to attend to the needs of marginalized populations? · What dimensions of diversity and what populations should be centered? · What goals are we trying to achieve in rethinking assessment? · In what specific ways might we make assessment more fair, equitable, and socially just? · How might opposition from within the field and from without be addressed

Discussant name
Discussant surname
Discussant affiliation

WEDNESDAY 3 JULY
Session 3.3 SYMPOSIUM/PANEL
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

57. Panelist for Organized Discussion: Rethinking Conceptions of Fairness and Equity in Assessment

Randy Bennett
ETS

Randy Bennett is Frederiksen Chair in Assessment Innovation at Educational Testing Service (USA). His recent focus has been on theorizing how assessment of, for, and as learning might be made more equitable. As past president of the International Association for Educational Assessment, and a collaborator with assessment organizations in several countries, he will bring to the organized discussion a perspective that tries to consider some of the impacts resulting from international population shifts, as well as other drivers of population diversity, to the challenge of rethinking assessment.

WEDNESDAY 3 JULY
Session 3.3 SYMPOSIUM/PANEL
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

58. Panelist for Organized Discussion: Rethinking Conceptions of Fairness and Equity in Assessment

Maria A. Ruiz-Primo
Stanford University

Dr. Ruiz-Primo is Associate Professor in the Graduate School of Education at Stanford University (USA). She was educated in Mexico through university level and brings extensive experience conducting educational assessment, evaluation, and research in the United States. Dr. Ruiz-Primo's work centers on classroom formative assessment, including understanding and developing teachers' assessment practices. The perspective she will bring to the symposium is informed by extensive experience working with teachers on issues related to assessment in the classroom with diverse populations.

WEDNESDAY 3 JULY
Session 3.3 SYMPOSIUM/PANEL
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

59. Panelist for Organized Discussion: Rethinking Conceptions of Fairness and Equity in Assessment

Jennifer Randall
University of Michigan

Dr. Randall is the Dunn Family Professor of Psychometrics and Test Development at the University of Michigan and the founding President of the Center for Measurement Justice (USA). Her work seeks to disrupt white supremacist, racist logics in assessment through justice-oriented practices that are explicitly and unapologetically antiracist. The perspective she will bring to the organized discussion will center Black, Brown, and Indigenous populations in the US and in other countries, exploring how the sociocultural identities of students can be deliberately considered and valued in the planning and development phases of assessment.

WEDNESDAY 3 JULY
Session 3.3 SYMPOSIUM/PANEL
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

60. Panelist for Organized Discussion: Rethinking Conceptions of Fairness and Equity in Assessment

Ye Tong
National Board of Medical Examiners

Dr. Tong is Senior Vice President at the National Board of Medical Examiners (USA). She was educated in the People's Republic of China through the university level. As Senior Vice President for Assessment Operations at the National Board of Medical Examiners, she has been spearheading efforts with partner organizations to reconsider approaches to assessment in medical education that better account for equity and fairness. As a former Vice President with Pearson, she also has had extensive experience with K-12 accountability testing. Shaped by her education and varied experiences managing high-stakes assessment programs, she will bring to the discussion a perspective that attempts to balance the often-times conflicting demands of clients, examinees, publics, measurement principles, and evolving conceptions of fairness.

WEDNESDAY 3 JULY
Session 3.4 SYMPOSIUM/ROUND TABLE
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

443. Adapting evaluation instruments to minority languages: Obstacles and alternatives

Nekane Balluerka, Jone Aliri1, Arantxa Gorostiaga
University of the Basque Country
Fréderique Vallar
Pearson Clinical & Talent Assessment España
Pablo Santamaria
Hogrefe TEA Ediciones
Maria E. Oliveri
University of Nebraska, Lincoln
Mirko Antino
The Spanish Journal of Psychology

Adapting evaluation instruments to minority languages is crucial to reduce the risk of cultural bias and to ensure that decisions based on test scores are fair for individuals from diverse populations. It is also necessary for their survival. Nevertheless, both researchers and test publishers face challenges that make such adaptations difficult. In the realm of research, beyond the difficulties that any process of adaptation to other languages may entail, there are additional challenges (e.g., obtaining a large sample of subjects to perform some type of psychometric analysis such as Item IRT) when the language is a minority language. Furthermore, it is difficult to obtain resources for adaptations to minority languages within competitive calls for research projects. Even if being successful in obtaining funding for such goal, it is difficult to disseminate the obtained results as many scientific journals refrain from publishing adaptations to minority languages due to their limited contribution and citation rates. As for test publishers, although they can respond to corporate social responsibility, adaptations to minority languages are not profitable. However, many governments in bilingual societies, where both majority and minority languages hold co-official status, have the need to assess psychological constructs in minority languages and thus are interested in funding the adaptation process. This roundtable will focus on adaptation to languages that have official status and are protected within the State to which they belong. The relevance of adaptation for the survival of minority languages will be addressed. Furthermore, the motivations and the obstacles faced by researchers, test publishers, and journal editors when making the decision to adapt or publish the adaptation of minority language assessment instruments. In addition, alternatives will be proposed to overcome these difficulties and to promote the adaptation of instruments to minority languages.

Discussant name: Nekane
Discussant surname: Balluerka
Discussant affiliation: University of the Basque Country

WEDNESDAY 3 JULY
Session 3.5 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

739. Computerized Adaptive Assessment of Cognitive and Non-Cognitive Competencies

María Dolores Nieto-Cañaveras
Nebrija University
Milagrosa Sánchez-Martín
Universidad Loyola Andalucía

The exploration of new procedures and instruments to measure cognitive and non-cognitive competencies in different applied fields such as education or working contexts is in vogue. In addition, it is common to measure both types of competencies simultaneously and/or together with other variables of interest, resulting in many cases in inefficient and time-consuming measurements due to the length of the measures used. In this scenario, Computerized Adaptive Testing (CAT) arises as a more efficient approach to assess both cognitive and non-cognitive abilities while significantly reducing testing time and producing equally or more precise estimates in comparison to traditional paper and pencil tests. This is the case with several international educative assessments, such as PISA, TIMSS, and PIRLS, and other emerging instruments to measure personality traits (e.g., entrepreneurial personality). Thus, this symposium aims to present various recent and ongoing applications of CATs to assess cognitive and non-cognitive competencies in different contexts. Specifically, the first and second presentations are focused on the process of item pool design, development, and calibration of two CATs to assess two types of cognitive competencies, spelling, and logical reasoning ability, in incoming university students. Special emphasis will be placed on addressing common challenges in this early and crucial phase of CAT development. The third presentation evaluates to what extent the use of optimal matching and a polytomous format can address the traditional limitations of forced-choice tests in adaptive personality assessment. Finally, the fourth presentation includes the evaluation of the performance of a CAT for measuring teamwork competencies in the professional context. In summary, the studies in this symposium aim to show different alternatives that CATs can offer in a variety of applied contexts.

Discussant name: José
Discussant surname: Muñiz
Discussant affiliation: Nebrija University

WEDNESDAY 3 JULY
Session 3.5 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

740. Development of a Forced-Choice Item Bank for Adaptive Personality Assessment: A Pilot Study

In recent decades, forced-choice (FC) tests have become established as an alternative to the traditional Likert format to minimize the effects of social desirability. However, developing a binary FC test, i.e., choosing between two options, is challenging due to the dichotomous format of the blocks, leading to lower reliability, and the ipsativity issues inherent to the FC format. Recently, the use of graded response FC format has been suggested, and solutions to the ipsativity problem have been proposed, based on optimal block matching. The goal of the current study is to evaluate to what extent the use of optimal matching and a polytomous format can address the traditional limitations of forced-choice tests. The study compares the psychometric properties of binary FC tests and five-category graded response tests, varying in the use of heteropolar blocks. Using a Likert-format personality item bank, rated for social desirability, two FC versions were created: one with exclusively homopolar blocks and another with 33% heteropolar blocks, optimally paired according to Kreitchmann et al.'s (2022) genetic algorithm. Both versions were applied in binary and graded response formats. The tests (heteropolar-binary, homopolar-binary, heteropolar-graded, homopolar-graded) were compared in terms of marginal reliability (according to Item Response Theory estimates), correlations between dimensions, and evidence of convergent and discriminant validity, focusing on the degree of ipsativity of the scores. Finally, recommendations are offered for constructing forced-choice tests.

Francisco J. Abad, Diego F. Graña
Universidad Autónoma de Madrid/España
Rodrigo S. Kreitchmann
Universidad Nacional de Educación a Distancia/España
Luis E. Garrido
Pontificia Universidad Católica Madre y Maestra/República Dominicana
Miguel A. Sorrel
Universidad Autónoma de Madrid/España

WEDNESDAY 3 JULY
Session 3.5 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

743. Multicultural Adaptive Assessment of Teamwork Competencies

María Dolores Nieto-Cañaveras
Nebrija University/España
David Aguado
Universidad Autónoma de Madrid/España
Ramón Rico
Universidad Carlos III de Madrid/España
Xie Xiao-Yun
Zhejiang University/China
Eduardo Salas
Rice University/EEUU

The assessment of teamwork competencies has shown to be of primary importance for team effectiveness. The current state of globalization and accelerated pace of change demand reliable and valid measures that can be efficiently used to obtain comparable score across different cultures. It is therefore crucial to have measures that allow comparisons between culturally different samples while considering invariant but also diverging aspects across countries due to culture. Thus, the purpose of this study was to develop and analyze the psychometric properties of a Computerized Adaptive Test (CAT) for measuring teamwork competencies in a multicultural team framework. The item bank was first developed in Spanish and then translated-backtranslated into Chinese and English, and administered to a large sample of respondents from three countries (Spain, China, and the United States of America). The unidimensional graded response model was used for parameter estimation in each dimension and subsample accounting for differential item functioning across countries. A post-hoc simulation study was carried out to assess the performance of unidimensional CATs to measure teamwork competencies. The results showed that the CAT provided efficient estimates, importantly reducing testing time. Practical implications of the study are finally discussed.

WEDNESDAY 3 JULY
Session 3.5 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

749. Development and calibration of a new item pool to measure logical reasoning ability in undergraduate students

Milagrosa Sánchez-Martín, Juan F. Luesia, Juan F. Plaza
Universidad Loyola Andalucía
María Dolores Nieto-Cañaveras
Nebrija University

Logical reasoning is a key construct within the paradigm of 21st-century skills, allowing effective reasoning in problem-solving. In higher education, this competence has proved to be especially relevant since it is necessary to solve academic tasks involving critical thinking, analytical thinking, and problem-solving, being measured in admission processes. Traditionally, logical reasoning instruments have been developed within a classical framework, with a fixed-length and the same items for all respondents. This implies certain vulnerabilities (e.g., cheating due to prior knowledge of item content) in high-stake contexts such as the one mentioned above. In this scenario, computerized adaptive testing (CAT) is an alternative that can solve some of the limitations of classical assessment. Thus, the purpose of this study was to design and develop a new item pool to be the core to measure logical reasoning in incoming students through CAT. This presentation shows the first steps regarding the construction of the initial item pool, which was administered to a sample of undergraduate students. The initial item pool was built and applied to a sample of undergraduate students. Preliminary analyses were conducted to select the items with better psychometric properties to be calibrated according to a unidimensional three-parameter logistic model. Subsequently, a post-hoc simulation study was carried out to assess the performance of the CAT and determine the features of the most optimal adaptive algorithm. The estimated unidimensional model is expected to show a good fit, preserving at least half of the original test items after conducting preliminary analyses of their properties. The CAT is also expected to be efficient in measuring logical reasoning ability with high accuracy while using a rather small number of items in the pool. Finally, some suggestions are offered regarding the different phases involving the construction of the item pool in this context.

WEDNESDAY 3 JULY
Session 3.5 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

751. Spelling competence in incoming university students: Creation of item pool based on current uses of Spanish spelling

Juan F. Luesia
Universidad Loyola Andalucía
María Dolores Nieto-Cañaveras
Nebrija University
Juan F. Plaza, Milagrosa Sánchez-Martín
Universidad Loyola Andalucía

Spelling is a fundamental aspect of written communication and is crucial to academic performance. In Spain, recent results have shown low scores in reading competence and literacy, which is coherent with the fact that teaching spelling has become a challenge in the international education setting. Acquiring and promoting appropriate spelling competence in university education must become a fundamental curriculum objective. The impact on the university stage students seems clear, and there is a growing concern for students' performance in academic tasks in which spelling mistakes are gradually worsening. Therefore, the assessment of this competence requires adaptation to new educational contexts. The development of Computerized Adaptive Tests (CAT) allows for a more efficient and individualized administration and scoring, but constructing such instruments implies a complex process that requires careful consideration. This study aimed to develop an item pool to assess spelling competence in new university students through a CAT. For this purpose, a map of spelling uses in Spain has been used as a starting point, which includes four main areas: use of letters, capital and lowercase letters, prefixation and composition, and accentuation. In addition, three categories of difficulty have been established: low, medium, and high. The initial bank was piloted on a sample of incoming students, expecting at least half of the initial items to show better properties for adaptive use. After that, a post-hoc simulation study was conducted to assess the performance of the CAT and determine the specifications of the most optimal adaptive algorithm. The developed CAT is expected to assess spelling competence efficiently, i.e., with an adequate accuracy from a reduced number of items. Finally, we will provide suggestions and future recommendations to facilitate the development and implementation of the CAT in computer-based higher education systems.

WEDNESDAY 3 JULY
Session 3.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

150. Development and Validation of Automated Video Interview Competency Assessments in Spanish

Gema Ruiz de Huydobro, Joshua Liff, Adam Bradshaw
HireVue

Using machine learning (ML)-based algorithms to score competency based asynchronous, video-interviews (AVIs), also known as automated video interview competency assessments (AVI-CAs), has proven to minimize human inconsistencies, and increase reliability and validity of structured interviewing. Despite the value derived from this assessment modality, it has been mostly productized as a solution applicable for English speaking populations, which limits the opportunities for non English speaking organizations to leverage its benefits. Meanwhile, in the HR Tech space there has been a heightened recognition of the need to consider an international angle in the development of AI-based solutions. Since Spanish is one of the most spoken languages globally, the current study's goal was to develop ML-based algorithms to score competency-relevant behaviors in AVIs in this language. ML techniques were applied to create algorithms to score four competencies, based on trained evaluator ratings. A sample of 27,033 unique AVI responses to past behavioral and situational questions were used to generate four interview scoring algorithms. Applicant verbal responses were extracted (i.e. transcribed and applied meaning through NLP approaches) from the AVI responses to use as the predictors in the machine learning analysis, with the criterion being the human evaluator ratings of the target competency. The four interview scoring algorithms were found to effectively replicate human evaluator ratings of AVIs with Multiple Rs ranging from 0.58 to 0.70. An average divergent validity correlation of .50 (N=17,044) among competencies was obtained. This research is an important step in demonstrating the effectiveness of AVI-CAs as a selection procedure in Spanish speaking employment contexts. Future research and partnership to increase data representativeness (i.e. country, industry, job level) is still needed to improve its applicability across more diverse populations.

WEDNESDAY 3 JULY
Session 3.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

305. Parameterizing Linear Logistic Test Model as a Neural Network

Denis Federiakin
Johannes Gutenberg University of Mainz
Lidia Dobria
University of Illinois at Chicago
Olga Zlatkin-Troitschanskaia3
Johannes Gutenberg University of Mainz

Recently, a new approach to estimating IRT model parameters via backpropagation has been suggested. This method is based on parameterizing IRT models as Artificial Neural Networks (ANNs). We introduce this logic of model estimation by describing a neuron, the building block of an ANN, as analogous to a regression model. Then, we describe autoencoder architectures of ANNs as analogous to PCA, since they condense information from multidimensional observations to a lower-dimensional representation. Then, we describe variational autoencoders, which learn latent variables for the data, similarly to IRT models. However, to preserve interpretability – one of the main advantages of IRT models – variational autoencoders should have shallow decoders. Employing a decoder that is only one neuron in depth allows item scores to be directly regressed from the latent variable, making the model equivalent to an IRT model. However, estimating Rasch-type models requires introducing specific constraints in the model, which have not been previously described. Particularly, in the output layer (estimating item parameters), weights (item discriminations) need to be initialized at 1 and not updated during the model training. Additionally, there is a necessity for an extra "input" layer (of a single neuron) in the decoder, that estimates the sample standard deviation. The parameterization of the Linear Logistic Test Model (LLTM) is even more complex. Since LLTM treats the parameters of 'cognitive operations' as precursors to item difficulties, in ANN parameterization, LLTM requires an additional layer in the decoder, which encodes the used Q-matrix. Using a real data example of a set of economics items aligned with the German higher education curriculum, we estimate the parameters of content-area difficulties in LLTM. We compare the ANN-based and the traditional parameterizations of LLTM and demonstrate that ANNs allow for nearly perfect recovery of IRT model parameters.

WEDNESDAY 3 JULY
Session 3.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

320. Machine Learning Unveils Factors Influencing Students' Math Performance Globally: Insights from PISA 2018

Liu Liu
University of Washington
Rui Dau
Arizona State University

Previous studies have established socioeconomic status as a key factor in mathematics achievement within the economic context (OECD, 2019). Research utilizing the Programme for International Student Assessment (PISA) data, such as Hanushek and Woessmann (2011) and Wang and Lin (2009), often depends on traditional statistical methods with a limited geographic focus. There remains a prominent gap in applying machine learning (ML) approaches for broader, cross-national comparisons in this field. This study leverages advanced ML techniques—including multiple linear regression (MLR), random forest (RF), extreme gradient boosting (XGBoost), CATBoost, and artificial neural network (ANN)—to analyze student math performance across the US, Chinese Taipei, Finland, and Argentina, utilizing PISA 2018 data. Addressing issues like missing data and multicollinearity, it standardizes a dataset featuring student and school characteristics across different education systems and controls for 24 diverse indicators such as SES, resources, parental involvement, and student self-efficacy to explore the complex factors impacting achievement. Preliminary findings indicate that CATBoost performs best in predictive accuracy. SES, home book count, and weekly math study time are the top keys affecting student performance in all the countries and regions studied. These factors suggest that access to resources and study habits are crucial, especially for students from lower SES. A notable discovery is the disproportionately high impact of owning 0-10 books at home, highlighting a critical threshold. This insight underscores policies should not only provide access to resources but guide students toward innovative learning strategies that encourage a breadth of interests and skill development. Contrastingly, the highest education level of parents is less influential. In sum, this study underscores the importance of data-driven policy recommendations to improve mathematics education.

WEDNESDAY 3 JULY
Session 3.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

484. Differences in Perceptions of Artificial Intelligence (AI) - powered Assessments and Impact on Test Performance

Justine Chalifour
HireVue

As Artificial Intelligence (AI) is increasingly used to preselect applicants, research has found that candidates tend to be more skeptical of these interviews compared to traditional assessments. In this AI era, where solutions are being designed and implemented globally, merely understanding candidate perception differences between both is insufficient. It is important to consider whether such differences in applicant reactions exist across demographic groups (e.g. age, gender, geographic region) to improve candidate experience, which also plays a role in candidate assessment performance. Assessment vendors, who open accessibility to job opportunities across a wider pool of diverse candidates, are in a great position to further understand this topic. The current study analyzes differences in applicants' satisfaction across demographics groups, as well as its impact on assessment performance. Applicants' overall satisfaction with an AI scored competency based AVI (asynchronous video interview) process (on a scale of 1-5; 1 =extremely dissatisfied and 5= extremely satisfied) was evaluated across age (n=29,977), gender (n=29,977) and region (n=50,154). Preliminary results from t-test and one-way ANOVA analyses indicate significant differences in mean satisfaction scores for age (Below 40: M=4.40, SD= .85; Above 40: M=4.30, SD=.89), gender (Male: M=4.31, SD= .90; Female M=4.45, SD=.81), and region (Africa: M=4.49, SD= .80; Asia: M=4.36, SD=.84 ; Europe: M=4.28, SD=.89 ; Latin America: M=4.54, SD= .70; North America: M=4.40, SD=.87; Oceania: M=4.30, SD=.85). Results from a multi-group Structural Equation Modeling (SEM) analysis evaluating group differences in the relationship between applicants satisfaction and assessment performance will be presented during the session. This research contributes to further understanding demographic differences in globalized personnel selection processes, where AI-based solutions continue to lead the way in the HR Tech space.

WEDNESDAY 3 JULY
Session 3.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

792. Predicting Missing Responses with Process Data in Large-Scale Computational Thinking Assessment

Qiwei He, Shihong Zhou
Georgetown University

In large-scale educational assessments, missing responses can be challenging and impedes the process of accurate estimation. Removing missing values, replacing them with zero values or average values usually leads to bias because of insufficient consideration on data correlation structure. In this study, we applied machine learning approaches affiliated with fine-grained process data in problem-solving procedure to predict missing responses. The objective of this study is two-fold: (1) to investigate whether individual-level process data could make a good prediction on missing responses, and (2) to identify the most robust behavioral factors in missing response prediction. A total of 11,468 respondents who gave a full response to all nine items in a computational thinking (CT) module in the International Computer and Information Literacy Study (ICILS) 2018 cycle were used in the present study. The reason for choosing the group with full responses was to set gold standards in machine learning training for missing responses item by item. In the multiclass prediction, it was found that the random forest approach significantly outperformed the support vector machine method when controlling on the same predictors. The prediction accuracy on missing responses is within a range of 0.6 to 0.9 on a macro level. On a micro level, the missing values in the high-performance group were better predicted than those in the low-performance group, probably because less information could be captured in the low-performance group for their more frequent item-skipping or disengagement behaviors. Three predictors (i.e., time spent on the tasks, the length of code blocks, and the remaining time) are found the most influential for missing response prediction. This study demonstrates the predictivity of process data on missing responses, which could be used as an alternative way to give a better estimation for the missing value, thus enhance the accuracy in item calibration.

WEDNESDAY 3 JULY
Session 3.8
Topic: Quantitative, qualitative, and mixed validation methods

191. Local item dependence in the SweSAT – how common is it and what are the causes?

Per-Erik Lyrén, Inga Laukaitytė, Christina Wikström
Umeå University, Sweden

Background A central aspect of testing is that responses to items are locally independent. While this is an underlying assumption of item response theory (IRT), violations to local item independence (i.e., local item dependence, LID) may also cause problems for classically scored tests, in terms of inflated reliability estimates. LID is usually associated with testlet-based formats, but since the potential causes of LID go beyond a common stimulus (e.g., speededness), LID should be examined regardless of test format. As goes for validity issues in general, LID is particularly important to examine in high-stakes testing contexts. Objectives The purpose of this study was to examine LID in the SweSAT, a test for selection to higher education in Sweden. The research questions were: a) How common is LID among the regular test items in the SweSAT? b) Is LID more common among items within the testlets-based subtests than the other subtests? c) For item pairs that exhibit LID, what are the proposed causes of this LID? Sample We studied four different test forms: spring 2022 (two forms; n=28,881 and 28,138, respectively), autumn 2022 (one form; n=38,068), and spring 2023 (one form; n=57,935). Method We used two measures for detecting LID: linear partial correlations and Yen's Q3. Focus groups with item writers/test developers were used to assess the causes of LID. Preliminary results In general, the number of item pairs with LID in the test was small. The most severe instances of LID were observed in two of the testlet-based subtests. While content multidimensionality was identified as one cause of small LIDs, it was not straightforward to identify the causes of larger LIDs. Implications The LID data is very useful as feedback information to the item writers and test developers. First, it shows that they do a very good job at avoiding redundant items. Second, by identifying causes of LID, remedial actions can be taken to minimize the risk of LID in future test forms.

WEDNESDAY 3 JULY
Session 3.8
Topic: Quantitative, qualitative, and mixed validation methods

293. An Application of Differential Item Functioning to an Adolescent Assessment Adapted for Adult Learners

John Sabatini
University of Memphis
Eli Jones
University of Memphis
John Hollander
University of Memphis

Differential item functioning (DIF) tests can be useful in determining whether items are of sufficient quality to achieve a comparable interpretation when adapted to populations beyond the originally intended respondents. We explore the presence of DIF in items on a reading components battery administered to adult learners originally scaled for use with adolescents. We apply multiple DIF detection methods to explore the functioning of the items for adults (focal group) when compared to adolescents (reference group). Samples. Focal group: 705 adults; reference group of 833 adolescents grade 6- 12, drawn from a larger national study. All participants took an online, reading component battery of 5 subtests (alpha >.88 and IRT-based scale for each subtest). We focus on a subsample of 63 items across five subtests. Analysis. We applied both traditional and IRT DIF using the difR package. We first applied the Mantel-Haenszel (MH) procedure. We calculated ΔMH coefficients and applied the ETS Delta scale to classify negligible (|ΔMH| ≤ 1), moderate (1 < |ΔMH| ≤ 1.5), or large (|ΔMH| > 1.5) effects. We then applied a logistic regression approach, followed by a standardization approach to test for items that consistently exhibited DIF across methods. All DIF tests were conducted using scale purification and the Benjamini-Hochberg p-value adjustment, to adjust for type I error rate inflation. Finally, we applied the likelihood ratio test using a 2PL model. Results/Discussion. Results were complex. When compared with the other two traditional methods, 27% items were flagged by DIF by at least two of the three traditional DIF methods, with only four items (6%) being flagged by all three methods. Of those flagged by 2 or more, 11 (17%) exhibited substantial DIF. However, neither group was favored consistently. The results revealed that most items were free from substantial DIF. Some subtest item groups performed better than others. We will discuss implications and IRT analyses.

WEDNESDAY 3 JULY
Session 3.8
Topic: Quantitative, qualitative, and mixed validation methods

528. Disentangling the Interplay of Emotional Intelligence, Personality Attributes, and MMIs in medical student selection

Mustafa Asil, Amy Bannatyne, Belinda Craig, Kirsty Forrest, Jessica Stokes-Parish, Jaclyn Szkwara
Faculty of Health Sciences & Medicine, Bond University, QLD, Australia

Introduction The selection of future medical professionals is a critical process that demands an assessment of not only academic ability but also essential non-cognitive traits. Emotional intelligence (EI), personality attributes and Multiple Mini Interviews (MMIs) have gained recognition as valuable indicators of a candidate's potential for success in medical practice. This study aims to explore the intricate relationship between emotional intelligence, personality attributes, and MMIs, all of which play pivotal roles in the selection of competent and empathetic healthcare professionals. Specifically, this study investigates whether: · There is evidence to support construct validity of MMIs, · Personality traits and EI predict candidates' MMI performance, · EI mediates the relationship between personality attributes and MMI performance. Method The individuals involved (N ≈ 800) are prospective students at Bond University's medical program for 2024. Psychometric testing and MMIs will take place in February. Data will be analysed using Structural Equation Modelling (SEM) to examine the complex relationships between EI, personality factors and MMI. The current literature lacks such a comprehensive method of analysis. Results This research study anticipates finding positive significant associations between emotional intelligence, specific personality attributes, and successful MMI outcomes. The SEM analysis will enable the estimation of direct and indirect effects, allowing for an exploration of the mediating role of EI. Discussion/Conclusion Through a comprehensive assessment of non-cognitive attributes, medical institutions can make informed decisions to nurture a diverse and competent healthcare workforce. The outcomes of this study hold the potential to guide the development of holistic admissions strategies, ultimately enhancing the selection of future medical professionals who can meet the evolving demands of the healthcare industry.

WEDNESDAY 3 JULY
Session 3.8
Topic: Quantitative, qualitative, and mixed validation methods

541. An Application of G-Theory and Many Faceted Rasch Measurement in Performance Assessment

Jon Twing1/1, Heather Klesch2/2, James Tognolini3/3
(1) University of Sydney, (2) Pearson Evaluation Systems, (3) University of Sydney

Local jurisdictions are struggling with fundamental aspects of measurement, concepts of reliability and validity, when attempting to assess these complicated constructs. Even when the constructs are simple, if the assessment includes performance assessment tasks, or technology enhanced items in additional to more traditional measures, simple issues of assessment reliability become muddled. In this regard, Smith & Kulikowich (2004) point out that performance and portfolio assessments have broadened our notion of what is considered evaluation but also point out that many of these new assessment tasks are not used in conjunction with traditional classroom learning. They go on to investigate such performance-based tasks using generalizability theory (Brennan, 2001) and many-faceted Rasch model measurement (Linacre, 1989, 1996b). The goal of their investigation was to "disentangle" person variance, task variance and judgement variance (i.e., rater judgment of person performance on the tasks) to understand both the effects of the raters and tasks but to also get better measures of person performance. While the current study is not a replication of Smith & Kulikowich (2004), it uses its concepts and groundwork in a more modern context from the U.S.. This study applies g-theory and the many-faceted Rasch model to a performance assessment task in teacher certification. The data used in this study comes from a completely crossed design (candidates by rubrics by raters) calibrating 15 candidates who each responded to 15 different tasks or rubrics scored each by two independent raters. Results show that the majority of variance is associated with assessment tasks and not raters but that the severity of raters varies greatly relative to candidate performance and task difficulty despite rater training and calibration. The research is beneficial to international assessment audiences in describing how to access analyses via spreadsheet or open-source software like R, procedures to follow and interpretations that can be made.

WEDNESDAY 3 JULY
Session 3.8
Topic: Quantitative, qualitative, and mixed validation methods

693. Developing the Facilitators and Obstacles of Recovery Scale (FOR-S) using the Delphi method: Insights from mental health professionals and service users

Georgina Guilera, Maite Barrios, Estefania Guerrero
University of Barcelona, Spain
Hernán María Sampietro
ActivaMent Catalunya Associació, Spain
Juana Gómez-Benito
University of Barcelona, Spain

Public mental health policies increasingly emphasize a recovery-oriented approach, which integrates perspectives from service users and embraces a holistic view of recovery, focusing on a meaningful and satisfying life. Understanding the facilitators and obstacles of the recovery process enables to address the challenges faced by professionals and users of mental health services. This research aims to develop the Facilitators and Obstacles of Recovery Scale (FOR-S). The process began with two Delphi studies, a collaborative and iterative method that facilitates the systematic gathering and refinement of expert opinions to ensure the inclusion of relevant indicators in the scale. The first Delphi study, conducted from March to June 2022, engaged 78 mental health professionals and produced 17 agreed-upon key facilitators and 23 on obstacles in the recovery process. The second Delphi study, conducted between September and November 2022, involved 101 mental health service users, reaching consensus on 8 facilitators and 12 obstacles. Consolidating findings from both studies, the research team merged indicators, ensuring a comprehensive list without redundancy. For each of the 30 indicators, two items were crafted, considering their readability and comprehensibility for the Spanish population. In the second phase, aimed at gathering evidence based on content, a panel of 17 mental health professionals, 17 service users, and 6 psychometricians assessed the clarity and relevance of items, instructions, and response alternatives using a 4-point Likert scale. Researchers selected one item per indicator based on the item content validity index. Furthermore, the wording of 15 items (50%) was improved based on experts' suggestions. The Delphi method allowed for the effective integration of both professionals' and first-person perspectives, positioning the 30-item FOR-S as a promising self-administered scale suitable for both clinical and non-clinical settings.

WEDNESDAY 3 JULY
Session 3.9 SYMPOSIUM
Topic: Quantitative, qualitative, and mixed validation methods

125. College Admission in Chile: Inequity, Social Unrest, and More Testing

Sergio Araneda
Caveon Test Security

The Chilean college admission system is an interesting case study for many reasons. The Chilean system is one of the few in the world that relies exclusively on numerical scores from both standardized tests and high school GPA. Also, the system is centralized, matching students and university programs according to their mutual preferences. In 2013, an international review of the test "Prueba de Selección Universitaria" (PSU) outlined many areas for improvement, including a problem with the opportunity to learn for students from vocational tracks. In 2020, student protests against the PSU led to significant disruptions, effectively canceling the History test that year and reflecting long-standing criticisms of the test's role in exacerbating socioeconomic inequalities. Responding to the technical criticisms of the PSU, calls for reform, and student protests, a committee of representative universities and the government – the Comité de Acceso – introduced major changes to the admissions test: the PSU would be replaced by a new battery of tests called "Prueba de Acceso a la Educación Superior" (PAES). To enact this change gradually, the committee also introduced a transitional battery of tests –"Prueba de Transición" (PDT) – to be used in 2021-2022 before the full adoption of the PAES. This symposium examines the adoption of these reforms. The first presentation examines the changes from a political standpoint, analyzing the minutes of the meetings of the Comité de Acceso. The second presentation analyzes fairness indicators related to the PAES. The third presentation focuses on the predictive validity of the PAES, and the last presentation focuses on the examinees' experiences reported on TikTok. This multidisciplinary effort can inform attendees about the Chilean case and bring useful lessons for other countries facing similar challenges related to deep inequalities, social unrest, and the role of standardized testing for college admissions in that context.

Discussant name:
Monica
Discussant surname:
Silva
Discussant affiliation:
Pontifica Universidad Catolica de Chile

WEDNESDAY 3 JULY
Session 3.9 SYMPOSIUM
Topic: Quantitative, qualitative, and mixed validation methods

418. Standardized Testing and Social Equity: An Evaluation of Recent Changes in Chile's University Admissions (Identifying biases by qualitative or quantitative methods)

David Torres Irribarra, Veronica Santelices
Pontificia Universidad Católica de Chile

In recent years, there has been increasing awareness and scrutiny around the world regarding the extent to which standardized assessments provide results that are fair for members of all groups within a population. Fairness is expected from large scale standardized assessments, most specially in cases where these assessments are used with significant consequences for examinees. Chile has a longstanding tradition of relying on a centralized university admissions system that relies on large-scale testing. During the last three decades, multiple reports have pointed out to the need for increased fairness in the Chilean admissions tests, particularly among different socio-economic groups and secondary school tracks. Partly in response to these demands, the Chilean agency in charge of the admissions standardized test development and implementation has introduced revisions to the tests. The broad modification agenda ranged from test content to the scoring process. This study examines the consequences of those changes both from a content and psychometric perspective. The evolution of different fairness indicators for the period 2018 and 2023 will be examined. Examination of the success of these measures, or lack thereof, in increased fairness among socioeconomic groups´ test results is an important component in maintaining the legitimacy of a national testing program, like Chile´s. Empirical evidence of the implications of recent changes is necessary to satisfy the social demand for fairness in large scale assessments with important consequences on students´ lives.

WEDNESDAY 3 JULY
Session 3.9 SYMPOSIUM
Topic: Quantitative, qualitative, and mixed validation methods

277. New Insights for Assessing the Predictive Capacity of Selection Tests in a Heterogeneous University System (Psychometric modeling)

Eduardo Alarcón-Bustamante, David Torres Irribarra
Pontificia Universidad Católica de Chile

One objective in education measurement is to evaluate the predictive capacity of a test in a population that is partitioned into groups. In general, the predictive validity is assessed by group, and the coefficients are compared among them. However, a trend might become apparent when individual examination of distinct groups is conducted, and it could potentially disappear when these groups are combined. The applicants to the university system who take the test do not have the same characteristics (e.g., sex, socioeconomic status, type of high school). Similarly, the universities that select the applicants have different characteristics (e.g., public or private, the different undergraduate programs they offer). In consequence, using a model for globally assessing the predictive capacity considering these characteristics. Given this heterogeneity, the relation between the admission test scores and any variable of interest is unlikely to be stable across all combinations and interactions of applicant and institution characteristics. Thus, we argue that characterising the prediction that test scores have over any variable of interest only with a single coefficient is an insufficient solution to adequately quantify and understand the potential variation of the predictive quality of the tests. We use the Law of Total Probability to learn about the predictive capacity of the selection test in the presence of groups. Through the marginal effect, we show how the heterogeneity of the groups affects the interpretation of the effect of the scores over an external variable to the test. Such interpretation shows the effect of the scores by considering differences in predicted outcomes and the size of each group. This effect is accordingly defined as a non-constant function of the scores (a decreasing one in this case). We show the results using a Chilean dataset. This fact reflects that the effect of the test scores is lower for higher scores and higher for lower scores.

WEDNESDAY 3 JULY
Session 3.9 SYMPOSIUM
Topic: Quantitative, qualitative, and mixed validation methods

387. Studying Examinees' Experiences Shared on Tik Tok about Standardized Testing and College Admission in Chile (Quantitative, qualitative, and mixed validation methods)

Xaviera Gonzalez-Wegener1/1, Sergio Araneda2/2
(1) Keele University, (2) Caveon Test Security

In this paper, we analyze the experiences of examinees taking the PAES tests in Chile (Prueba de Acceso a la Educacion Superior), a new test battery used for college admission, first implemented in November 2022. Our analysis centers on a sample of TikTok videos posted by Chilean test-takers preparing for PAES. Utilizing Araneda et al.'s conceptual framework for test-taking experiences (2023), we categorize interactions between examinees and the test. This approach allows us to formulate hypotheses about potential invalidities based on the topics discussed in the TikTok videos; using the experiential approach to validation (Araneda & Sireci, 2023). This study is significant as it utilizes TikTok, a novel information source, to gain insights for test developers. This approach is instrumental in guiding validation efforts and influencing test design decisions. Additionally, it offers insights into the unique challenges faced by the new PAES test battery, especially in its inaugural years. The study also contextualizes the recent historical events in Chile, highlighting how the analysis of student experiences can inform future research on the PAES tests. Our methodology involves analyzing a sample of 40 TikTok videos flagged as related to PAES test experiences. We will include video transcripts and comments, with two raters identifying types of interactions, sense-making stages, and employing Grounded Theory for topic detection. These identified topics will be used to pinpoint potential invalidity sources in the PAES tests. The results of our study will be discussed in the context of the evolving field of educational assessment and the role of social media in understanding user experiences. We will reflect on the implications of our findings for test developers and policymakers, especially in the dynamic educational landscape of Chile.

WEDNESDAY 3 JULY
Session 3.9 SYMPOSIUM
Topic: Quantitative, qualitative, and mixed validation methods

389. Change is never easy: the case of Chilean College admission system and their new battery of standardized tests PAES (Quantitative, qualitative, and mixed validation methods)

Fernanda Gandara1/1, Sergio Araneda2/2
(1) Room to Read, (2) Caveon Test Security

In this presentation, we explore the complex transition from the old PSU test battery to the new PAES system in Chile by analyzing the minutes of a committee of representative universities and the government – Comité de Acceso – that led this change. This presentation aims to examine the political landscape that led to and fueled the changes to the current Chilean admissions system. By analyzing official documents and records from the 2019-2023 period, including the official transcripts of the Comité de Acceso meetings, we will synthesize the main themes and challenges encountered. We will contrast these themes and challenges with the shortcomings identified by previous reports and by the public. Ultimately, we will aim to uncover the rationale behind technical and non-technical decisions surrounding the changes to the Chilean admission tests and the governing logics that prevent or enable certain types of improvements. The results of this study should help understand the priorities and limitations of the current set of changes. The study's results will provide insight into the legal and structural conditions that may enable or prevent change. Last, by contrasting the language in the official documents with the narratives encountered in other spaces (e.g., student organizations), this work will evidence the complexities of making changes to college admissions. Our work will support other policymakers considering similar changes to their admissions systems by identifying the obvious and subtle challenges inherent to reforms transitioning from one battery of standardized tests to a new one.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

564. Validation of the Maternal Postpartum Stress Scale (MPSS) in Spanish population: analysis of the internal structure.

Alejandro de la Torre-Luque
Universidad Complutense de Madrid
Adrian Ruiz-Perete
Universidad Complutense de Madrid. Spain
Sergio Martinez-Vazquez
Universidad de Jaén, Spain
Rafael Caparros-Gonzalez
Universidad de Granada, Spain

Although scales that evaluate postpartum stress exist, they lack specificity, and most are universal. The Maternal Postpartum Stress Scale (MPSS) scale was created because there is a need to assess maternal stress during postpartum maternity leave. The introduction of MPSS has enriched the evaluation tools for postpartum stress and has helped understand maternal stress at various postpartum time points and identify women at high risk for postpartum stress, which may be beneficial for formulating personalized interventions. The scale also measures stress during the first postpartum year empowering its screening application. The total scale and three subscales can be calculated separately to provide detailed information about stressors that mothers struggle with. The aim of this study is to validate the MPSS in the Spanish population by analyzing the internal structure of its translated version. The sample comprised 167 pregnant women (M=34.26 years, SD=4.71). Besides the MPSS, there was administered a demographic scale, as well as Beck Scale for Suicide Ideation (BSS) and Edinburgh Postnatal Depression Scale (EDPS). The MPSS data were analyzed. We first checked item communalities, observing 3 items with unsatisfactory levels (h2

WEDNESDAY 3 JULY
Poster Session 2
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

499. Evidence of content validity: Experts judgment in an Externalizing Problem Behavior Scale in Adults.

Lidia Torres Rosado, Cinta Mancheño Velasco, Andrea Blanc Molina, Manuel Sánchez García
Universidad de Huelva

Recent models such as the Alternative Model of Personality Disorders (AMPD-APA, 2013) and the Hierarchical Taxonomy of Psychopathology (HiTOP) have introduced a novel definition for Externalizing, addressing criticisms directed at existing taxonomies. The Externalizing Spectrum Inventory (ESI) (Krueger et al., 2007) is currently the most comprehensive tool for measuring Externalizing, developed within the framework of the HiTOP model. However, a notable limitation of this instrument is that it only comprehensively covers one dimension of Externalizing. This may be attributed to the fact that the instrument's development predates the publication of the HiTOP model. Hence, we propose the development of an Externalizing Behavior Problems Scale for Adults (EPCEA), designed to align with both DSM and HiTOP models. Firstly, the Externalizing construct was defined from an integrative perspective (Torres-Rosado L. et al., 2023). Subsequently, a detailed specification and an extensive item bank (four times the specified number of items) were created. Finally, for content validity, items underwent evaluation for relevance by 38 experts, employing three indices: Content Validity Index, Validity Coefficient (Lawshe, 1975) and Aiken's V (Aiken, 1980). Experts were also encouraged to provide suggestions and comments for each item. The outcome is presented in a detailed specifications table, delineating dimensions, subdimensions, weights, and observable indicators. Additionally, expert judgments' results for disorders falling under Externalizing are outlined, encompassing Antisocial, Borderline, Paranoid, Histrionic and Narcissistic personality disorders, as well as Attention Deficit Hyperactivity Disorder (ADHD). The majority of disorders and facets exhibited favorable relevance indices. In instances where relevance was lacking, items were redefined based on expert feedback

WEDNESDAY 3 JULY
Poster Session 2
Topic: Quantitative, qualitative, and mixed validation methods

580. Are Open-Ended Demographic and Non-Demographic Items Useful in Evaluating Data Quality? Examining Responses from Adults Recruited via Amazon's MTurk.

Alexis D. Webster, Anita M. Hubley
University of British Columbia

Framework: Many researchers rely on online platforms such as Mechanical Turk (MTurk) for data collection as large datasets can be acquired quickly and conveniently. In recent years, however, such datasets have shown high proportions of low quality responses, which can distort reliability, validity, and other study results in unforeseen ways. There are numerous data quality control methods, but they tend to be underused and underreported and many also require further study to better understand their utility. Objective: The purpose of the present analysis was to examine the prevalence of low quality responses to four different open-ended items (2 demographic, 1 general, and 1 study-specific items) and identify their problematic characteristics. Methodology: Participants were recruited using MTurk and completed a variety of self-esteem, narcissism, health, demographic, and data quality check items online in March 2023. Duplicate responses and possible bots were removed. Sample: The final sample consisted of 343 adults ages 21 to 72 (43.7% women and 56.3% men). Results: We found that 86% of participants provided at least one low quality response. Problematic responses were more prevalent in the two non-demographic items that required longer responses (84.8% and 47.8%) than the demographic items (27.1% and 17.5%). The most common problematic characteristic was a response that was identical or uncannily similar to others' distinct responses. Other common problems were: responses that were irrelevant, plagiarized, or vague; known low quality; and spelling/grammar errors. Implications: Open-ended items can provide rich information for assessing data quality. We provide a set of recommendations for the number and type of open-ended items to use as well as criteria to exclude respondents. Online data collection requires researchers to be vigilant; use of a variety of effective data quality checks can increase confidence in research findings.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Quantitative, qualitative, and mixed validation methods

578. Response Processes Validity Evidence Using Cognitive Interviews: Clear Reporting Necessary to Ensure Good Measurement Practice.

Amanda Rose Dumoulin, Sophie Ma Zhu, Anita M. Hubley
University of British Columbia

Framework: The Standards for Educational and Psychological Testing (AERA et al., 2014) present response processes (RP) as an important source of validity evidence that examines the fit between what respondents actually do and what is expected of them. The most common method of examining RP evidence is cognitive interviews (CIs). Objective: The purpose of this scoping review was to examine how CIs were used to evaluate RP as a source of validity evidence. Methodology: Using medical education as an example, we searched six databases for works published from 2018 to 2023 using "response process*", "valid*", and "cognitive interview*" as search terms and focused on samples of physician trainees—medical students (29%) or residents (71%). Sample: This search resulted in 16 studies that used CIs in the development and/or validation process of self-report (75%) or assessment (25%) measures. Results: Ten studies involved the development of new scales, 4 were revising existing measures, and 2 were examining validity evidence for existing measures. CIs were frequently used in pretesting to identify clarity and feasibility issues or gather general feedback. All but one study reported other sources of validity evidence (most frequently test content and internal structure) in addition to RP. Often, RP evidence was not clearly differentiated from test content evidence. None of the studies described a priori the intended understanding of the items to which they could compare the RP data collected and establish an argument for the proposed interpretation of scores. Implications: Due to the lack of detailed information in the majority of the studies, (a) it is unclear if many of the authors had an adequate understanding of how to use CIs to provide RP evidence for validity, and (b) procedurally, it would be impossible to replicate the studies. We provide several recommendations to improve both research practice and the clarity of reporting of RP validity evidence when CIs are used.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Quantitative, qualitative, and mixed validation methods

738. Identifying Imbalances and Gaps in Psychometric Evidence: A Reliability and Validation Synthesis of the Rosenberg Self-Esteem Scale.

Sophie Ma Zhu, Robert J. Ruddell, Anita M. Hubley
University of British Columbia

Framework: Self-esteem is one of the most investigated social sciences constructs and the Rosenberg Self-Esteem Scale (RSES) is the most widely used measure of self-esteem. To be confident in using the RSES, it is important to examine the breadth and quality of its reliability and validity evidence. Objective: The purpose of this study was to review the reliability and validation practices used to evaluate the unmodified English RSES as guided by the 2014 Standards for Educational and Psychological Testing. Methodology/Sample: We conducted a search using the scale name and 26 search terms in 6 social & medical science databases; 30 studies met our inclusion criteria. We recorded types of reliability & sources of validity evidence reported plus related methodological details. Results: Only one (77%) or two (23%) sources of validity evidence were presented in studies. The vast majority (83%) examined internal structure, often testing multiple models. This evidence provides important information about number of scores needed to best capture variance in item responses, but it contributes relatively little to our understanding of score meaning. Reliability, consisting exclusively of internal consistency estimates, was reported in 80% of studies and was consistently satisfactory only for a total score. Nine studies (30%) examined relations to other variables evidence but tended to be poorly reported. The rationale for choice of measures and clarity in the interpretation of validity evidence was generally lacking. Few RSES validation studies have examined response processes (1), consequences of testing (1), or test content (0). Consequently, evidence to support the validity of inferences made from RSES scores is surprisingly limited. Implications: This study illustrates the utility of a reliability and validation synthesis in (a) highlighting the presence of imbalanced psychometric evidence for a measure and (b) identifying specific evidential gaps to guide future efforts.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Testing equivalence by psychometrics methods

530. Examining generational test score changes in Spatial and Word Analogy performance: Insights from Austrian conscript data.

Alina Bugelnig
Military Psychological Service, Austrian Federal Ministry of Defence; Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna
Maria Gruber, Alexander Birner
Military Psychological Service, Austrian Federal Ministry of Defence
Jakob Pietschnig3/4
Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna

During the 20th century, there has been a positive trend in generational IQ test score changes (the Flynn effect). However, over the past few decades, the IQ change trajectory seems to have become more erratic. In this study, data from Austrian conscripts are used to investigate the Flynn effect in two measurement-invariant tests from 2011 to 2021. Conscript samples prove particularly valuable in exploring such changes because they are by default population representative for young men. Firstly, Test score changes were examined in a Spatial Visualization test (N = 76,064), revealing an increase of 3.41 IQ points per decade. This contrasts evidence from recent studies in Germanophones that indicated a negative Flynn effect from 1977 to 2014. Changes in the Word Analogy test (N = 77,812) were virtually zero, revealing an increase of mere 0.19 IQ points per decade. The results support the growing body of evidence pointing to increasingly inconsistent and domain-specific patterns of the Flynn effect in recent decades. Conceivably, these results could be attributed to an increasing differentiation in cognitive abilities within the general population.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Testing equivalence by psychometrics methods

551. The English Version of the Satisfaction With Life Scale (SWLS) for Asian International Students in the United States: A Cross-Cultural Study.

Giusy Danila Valenti1/1, Palmira Faraci2/2
(1) University of Palermo, (2) University of Enna "Kore"

The Satisfaction With Life Scale (SWLS) is a popular instrument for assessing the cognitive component of global subjective well-being and has a long history of cross-validation studies examinations. However, the psychometric properties of the SWLS have never been tested when used with non-native speakers. This study examined the factor structure, internal consistency and measurement equivalence of the English version of the scale when used with non-native English speakers. The total sample consisted of 338 individuals, of whom 167 were Asian international university students living in the United States (50.3% females; Mage = 23.82, SD = 3.78) and 171 were Italian university students living in Italy (69.6% females; Mage = 22.38, SD = 4.24). The dimensionality of the scale was tested by a Confirmatory Factor Analysis (CFA) and a Multi-Group Confirmatory Factor Analysis (MG-CFA) to examine measurement invariance. The results confirmed the unidimensionality of the scale [χ^2 = 9.815; df = 5; CFI = .989; TLI = .977; RMSEA = .076 (.000–.146); SRMR = .023; AIC = 2,560.463; BIC = 2,607.233; aBIC = 2,559.741], and the achievement of full strict invariance suggested that the SWLS items were similarly structured in both samples. Furthermore, the scale showed an adequate degree of internal reliability (α = .863, ω = .866). The present study supports the cross-validity of the English version of the SWLS and indicates its robustness and suitability for assessing life satifaction in non-native English speaker.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Testing equivalence by psychometrics methods

92. Factorial equivalence of the core cognitive abilities of the WAIS-IV across the US and UK.

Hannah Cruickshank Campbell, Christopher J. Wilson
Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, Australia
Abigail Batty
Research and Development. Pearson Clinical Assessment, London, United Kingdom
Stephen C. Bowden
Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, Australia

Factorial equivalence is the demonstration of the measurement invariance of the factor structure of an assessment across groups. When measurement invariance is found, the psychological meaning of the constructs underlying an assessment can be generalised to both groups. Measurement invariance must be demonstrated before any cross-cultural comparisons can be undertaken. This study explored the cross-cultural factorial equivalence of the factor structure underlining the Wechsler Adult Intelligence Scale–Fourth edition (WAIS-IV) using the ten core subtests. Raw score data from nationally representative samples of the US and UK were used as input data. Previous research has demonstrated the factorial equivalence across populations from the US and Canada on the WAIS-IV and the US and Australia on the WAIS-III. Results of the baseline model estimation using confirmation factor analysis found a previously established modified four-factor model displayed the best fit in both the US and UK samples independently. Using the modified four-factor model, tests of factorial invariance were explored in the recommend hierarchical sequence. Strict factorial invariance was observed across the US and UK samples on the WAIS-IV. These results support the use of the published four-factor model across different populations. Further, the results allow for the generalisability of convergent and discriminant validity using the WAIS-IV.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Testing equivalence by psychometrics methods

687. Comparing Item Parameters and Scales of Adolescent Assessment Adapted for Adult Learners.

John Sabatini
University of Memphis/USA

We compare and contrast methods of evaluating items and scaling assessments of reading skills tests between adolescent and adult learners. Sample: Item parameters and vertical scales were computed using item response theory (IRT-2PL) for a national sample of ~31,000 US students in grades 5-10 for a multiform, 6-subtest component skill battery: word recognition/ decoding (WRDC), vocabulary (VOC), morphology (MOR), sentence processing (SEN), reading efficiency (EFFIC), and reading comprehension (RC). Reliabilities for each subtest range from .65-.96. For this study, we collected four subgroup samples of at-risk readers in adult education or postsecondary settings (n=~1450). Analysis approach: We first used the IRT-2PL fixed parameters originally estimated for the adolescent sample to conduct a calibration analysis of the adult sample. We then created new item parameters and scale scores based on the adult sample only. We compared adult vs adolescent scale score means, correlations, and item fit statistics. Results: Results were complex. Intercorrelation among subtests within the adult scale scores showed similar patterns as one observes with the adolescent scales. Visual inspection of the slopes shows how the various adult subsamples differ somewhat in the slope of item difficulty change with ability from adolescent patterns. Adult scale score means were higher for WRDC and SEN, but lower for MOR and RC than adolescent scale means. Correlation between adult and adolescent scales ranged from r=.75 (RC) to .93 (SEN). Discussion: This research investigates quantitative and qualitative techniques for evaluating the quality and validity of tests designed for one population (adolescents) when applied to another (adults).

WEDNESDAY 3 JULY
Poster Session 2
Topic: Testing equivalence by psychometrics methods

663. Modelling indicator-specific effects in longitudinal invariance: the case of the Revised-University of California at Los Angeles Loneliness scale.

Laura Galiana, Irene Fernández, Sara Martínez-Gregorio, Adrián García-Molla, José M. Tomás
University of Valencia

Longitudinal measurement invariance is an extension of factor analysis aimed at establishing invariance across time. Method effects produced by repeatedly measuring the same indicators is a common problem whose traditional solution lies in correlating residual variances of equal indicators, an application of the correlated uniqueness model (CTCU). However, this model presents some limitations compared to the correlated traits and correlated methods (CTCM). We aim to compare these two models in a longitudinal invariance routine of the Revised-University of California at Los Angeles Loneliness scale (R-UCLA). Data from 6488 Spanish older adults from waves 5, 6, 7, and 8 of the Survey of Health, Aging and Retirement in Europe (SHARE) was used. Measurement equivalence was tested via confirmatory factor analysis. Two sets of hierarchical increasingly restricted models were employed to assess invariance: one modeling a CTCU model and the other one with a CTCM model. Comparison was done using Comparative Fit Index differences (ΔCFI), Root Mean Squared Error of Approximation differences (ΔRMSEA), Akaike Information Criterion (AIC), and Bayesian Information Criteria (BIC), with smaller values indicating better fit. Robust maximum likelihood was used for the estimation in Mplus 8.9. With both methods, longitudinal measurement invariance was achieved, with adequate fit for the scalar model. When they were compared, subtle differences were found: ΔCFI favoring the CTCU model and ΔRMSEA favoring the CTCM model. Information criteria were better for the CTCM model. Although literature suggests that indicator-specific effects generalize beyond time points and modeling them with factors would be desirable, most research still employs correlated uniquenesses. This study suggests that using indicator-specific factors better addresses indicator specificity, balancing sensitivity and specificity. Future research comparing these models in terms of reliability and validity is welcomed.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Testing equivalence by psychometrics methods

231. Approximate invariance of ARDES measures in 7 countries using Alignment Analysis: Argentina, Australia, Brazil, China, Spain, UK, and USA.

Pablo Doncel, Cándida Castro
CIMCYC (Mind, Brain and Behaviour Research Centre), Faculty of Psychology, University of Granada, Spain
Rubén D. Ledesma, Silvana A. Montes
IPSIBAT, Instituto de Psicología Básica, Aplicada y Tecnología, CONICET (National Scientific and Technical Research Council) and Universidad Nacional de Mar del Plata, Argentina
D. Daniela Barragan
George Mason University, USA
Oscar Oviedo-Trespalacios
Delft University of Technology, The Netherlands, (5) UFPR (Federal University of Parana), Brazil
Alessandra Bianchi, Natalia Kauer
UFPR (Federal University of Parana), Brazil
Weina Qu
CAS Key Laboratory of Behavioural Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China
Jose-Luis Padilla
CIMCYC (Mind, Brain and Behaviour Research Centre), Faculty of Psychology, University of Granada, Spain

The Attention-Related Driving Errors Scale (ARDES) is a self-report measure of individual differences in driving inattention. ARDES was originally developed in Spanish (Argentina), and later adapted to other countries and languages. Evidence supporting the reliability and validity of ARDES scores has been obtained in various different countries. However, no study has been conducted to specifically examine the measurement invariance of ARDES measures across countries, thus limiting their comparability. Can different language versions of ARDES provide comparable measures across countries with different traffic regulations and cultural norms? To what extent might cultural differences prevent researchers from making valid inferences based on ARDES measures? Using Alignment Analysis, the present study assessed the approximate invariance of ARDES measures in seven countries: Argentina (n = 603), Australia (n = 378), Brazil (n = 220), China (n = 308). Spain (n = 310), UK (n = 298), and USA (n = 278). The three-factor structure of ARDES scores (differentiating driving errors occurring at Navigation, Manoeuvring and Control levels) was used as the target theoretical model. A fixed alignment analysis was conducted to examine approximate measurement invariance. 12.3 % of the intercepts and 0.8 % of the item-factor loadings were identified as non-invariant, averaging 8.6 % of non-invariance. Despite substantial differences among the countries, sample recruitment or representativeness, study results support resorting to ARDES measures to make comparisons across the country samples. Thus, the range of cultures, laws and collision risk across these 7 countries provides a demanding assessment for a cultural-free inattention while-driving. The alignment analysis results suggest that ARDES measures reach near equivalence among the countries in the study. We hope this study will serve as a basis for future cross-cultural research on driving inattention using ARDES.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Testing equivalence by psychometrics methods

462. A Psychometric Analysis of the CES-D 10 Using the South African National Income Dynamics Study (NIDS) Panel Data.

Richard Fletcher, Hermione Clayton, Natasha Bradley, Hayley Webb
Massey University

Objective: Using confirmatory factor analysis (CFA), to investigate the underlying factor structure of the CES-D-10 in a nationally representative panel data set within South Africa. Method: A total of N=12233 adult participants who identified as either African, Coloured African, Indian/Asian or White, were selected and assessed over five waves of data. CFA models were examined for each group, as well as a series of longitudinal invariance models for each of the five time points. Results: Reliability estimates for the 10 items across the groups over time were generally low. The two positively worded items presented with low item total correlations, however, when removed reliability increased across all groups to > 0.70. CFA results showed a similar pattern when the two positively worded items were removed. The model fit for the 8 items was high for African males and females, whereas the fit was low to acceptable within the remaining groups. Results for the invariance models showed only weak invariance for African males and females. Conclusions: Consideration should be given for the removal of the two positively worded items when estimating depression/wellbeing cut scores. Therefore, caution is warranted when using the CES-D-10 to monitor changes in depression/wellbeing in longitudinal studies.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Psychometric modeling

800. An empirical comparison of IRT-based and Generalizability Theory-based approaches for estimating Conditional Standard Errors of Measurement in personality testing.

Gempp René
UDP

While the Standards for Educational and Psychological Testing recommend reporting Conditional Standard Errors of Measurement (CSEM) for test scores, this issue is rarely discussed for personality tests. In fact, there are very few published studies on CSEM in personality measures despite their well-established use in educational and cognitive tests. One explanation is the general misunderstanding that CSEMs can only be estimated by applying IRT models, while most personality tests are based on the Classical Test Theory (CTT) paradigm. Additionally, a few simple methods for estimating CSEM within the CTT paradigm are valid for dichotomous rather than polytomous items typically found in personality questionnaires. However, the estimation of CSEMs is feasible through Generalizability Theory (GT), an ANOVA-based extension of the CTT. GT provides a computationally simple and efficient method for estimating the CSEMs for any test. This study empirically compares the simplified version of the GT approach for estimating CSEMs, suggested by Brennan (1998, 2001), with an IRT-based method. Both procedures were applied to a personality test of 44 items, each with five response options based on the five-factor model, answered by 949 job applicants. First, a one-facet GT model was applied to each of the five dimensions, and the relative CSEM was estimated using the simplified procedure developed by Brennan (1998, 2001). Then each big-five factor was calibrated using Samejima's Graded Response Model, and the CSEMs were estimated using the method described by Wang, Kolen, and Harris (2000). The results showed that the CSEMs obtained using the GT approach were comparable to those estimated using the IRT method. The practical implications for measurement practitioners are also discussed.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Psychometric modeling

796. Moderated Nonlinear Factor Analysis for Measurement Invariance.

Joseph Kush James
Madison University

Measurement invariance (MI) is a foundational property of highly valid measures. MI represents the degree to which observed item distributions are the same at various levels of the latent variable, regardless of group membership. Traditional multigroup analytic approaches to MI testing examine discrete groups tested in succession. Moderated nonlinear factor analysis (MNLFA) overcomes these two apparent limitations by considering the moderating effects of multiple covariates (categorical or continuous) simultaneously. After introducing the MNLFA model, a motivating example is explored in which multigroup analyses indicate no presence of differential item functioning (DIF) among items. However, a more nuanced approach utilizing the MNLFA model indicates concerning levels of DIF not detected by the traditional multigroup approach. Various implications for psychometricians and assessment professionals are further discussed.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Psychometric modeling

129. Matching student's vocational interests to study environments: Model-based mapping of university faculties in Holland's circumplex RIASEC space.

Lisa Bailey, Gideon de Bruin
Stellenbosch University
Brandon Morgan
University of Johannesburg

Holland's circumplex RIASEC model of vocational interests can be used to describe both persons and environments, which conveniently allows for the examination of the congruence between persons and environments. High levels of congruence between students' vocational interests and their chosen fields of study are likely to lead to better academic outcomes. However, finding model-based vocational interest profiles of study environments or fields of study is under-researched. Job classification systems, such as Holland's Dictionary of Occupational Titles, primarily focus on job environments rather than fields of study and may be sub-optimal in guiding students towards choices of degree programs. The present study aimed to locate eight university faculties (i.e. Arts, Agriculture, Education, Law, Economics and Management, Engineering, Medicine, and Science) as representatives of study environments, in Holland's two-dimensional RIASEC space. We used the responses of 2946 university students to the South African Interest Inventory. The structural summary method (SSM) was used to estimate the fit of a model-based circumplex interest profile for each of eight faculties and to locate them in the circular RIASEC space with respect to elevation, displacement and amplitude. Overall, results indicated good fit with the predictions of Holland's model. The angular locations of the faculties in the RIASEC space matched theoretical expectations, with Medicine and Agriculture being noticeable and interesting exceptions. Employing the SSM yielded faculty-specific environmental profiles against which student profiles can be compared. Overall, these results support the utility of Holland's vocational theory in an African context and represent a step towards more detailed model-based comparisons of students and their chosen study environments.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Psychometric modeling

178. Exploring the dimensionality of a neuropsychological test battery in a memory clinic setting using psychometric network analysis.

Mathilde Bastien
Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland,
Christian Chicherio
Geneva Memory Center, Department of Rehabilitation and Geriatrics, Geneva University Hospitals, Switzerland ; Center for Interdisciplinary Study of Gerontology and Vulnerability, University of Geneva, Switzerland
Salome Döll
Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland
Federica Ribaldi, Giovanni B. Frisoni
Geneva Memory Center, Department of Rehabilitation and Geriatrics, Geneva University Hospitals, Switzerland
Thierry Lecerf
Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland

Common neuropsychological batteries assess several cognitive abilities through standardized tests, with score interpretation relying on test manuals. However, psychometric properties of the scores depend on test and sample characteristics. Thorough investigations into score validity is crucial for clinical relevance, involving exploration of the relationship between neuropsychological measures and theoretical attributes. Factor analysis is classically used to examine the internal structure of the scores by identifying latent variables. A recent method, Exploratory Graph Analysis (EGA), part of the network psychometric approach, allows to investigate the dimensionality of psychological instruments. Instead of focusing solely on latent factors influencing observed variables, this approach considers direct connections among variables. It offers a valuable perspective for exploring the dimensionality of neuropsychological tests and understanding the intricate relationships between various components, contributing to a more comprehensive grasp of cognitive functioning. This research uses psychometric network analysis to examine the structure of neuropsychological tests administered at the Memory Center of Geneva. The study includes 212 healthy adults, 391 patients with subjective cognitive decline, and 677 patients with mild cognitive impairment. While EGA investigates the internal structure, bootEGA explores the reproducibility and the generalizability of EGA-identified dimensions. This study considers general demographic variables like age, gender, and education. EGA and bootEGA revealed 5 to 7 clusters, varying across groups. Test structures differ between groups. Despite cluster variability, some, like those with the Stroop and Digit Span tests, showed consistency across all groups. While the study validated the relevance of specific tests in clinical assessments, it also highlighted the complexity and potential variability in understanding certain cognitive measure.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Psychometric modeling

478. Rasch analysis of the 21 item Depression, Anxiety and Stress Scale in a mild traumatic brain injury sample.

Background: Psychological difficulties following mild traumatic brain injury mTBI are common and psychological distress after mTBI correlates with symptoms and recovery. The 21-item Depression, Anxiety and Stress scale DASS-21 has been used to measure distress after brain injury although some uncertainty remains as to the precise number of dimensions underpinning it. The present study used Rasch analysis to examine DASS-21 in a sample of people reporting mTBI. Method: Secondary analysis of data from 347 62% female individuals following mTBI in New Zealand. Rasch analysis was completed using RUMM 2030 software and included these steps: 1. Likelihood Ratio test; 2. Testing overall model and individual item fit by: Examining for disordered thresholds; Resolving local dependency by creation of super-items; Testing for unidimensionality and differential item functioning DIF. Results: Initially overall fit for the full 21 items was poor $\chi 2=240.74$ 105, p

Richard Siegert
Auckland University of Technology
Josh Faulkner
Victoria University of Wellington, New Zealand
Deborah Snell3/3
University of Otago, Christchurch, New Zealand

WEDNESDAY 3 JULY
Poster Session 2
Topic: Psychometric modeling

437. Measuring How Individuals Mentally Relate Science to Religion.

Rizqy Amelia Zein, Mario Gollwitzer
Ludwig Maximilian University of Munich

When trying to make sense of what is going on in the world and their personal lives, people often refer to scientific and religious explanations. Some scholars have argued that these two types of explanations contradict each other, yet evidence suggests that people think about the science-religion relationship in much more refined and integrative terms. Yet, a measure that captures the entire variety of people's mental representations of the science-religion relationship has been lacking so far. Based on the pertinent literature, we propose that these representations exist in five types. Therefore, we develop a measure that allows us to categorize people into one of these types and hypothesize that participants' responses reflect maximum agreement (i.e., unfolding) rather than maximum performance (i.e., dominance) response behavior. Furthermore, we also hypothesize that these types can be ordered along a single continuum of conflict-compatibility. To test this, we created an item pool derived from extensive evidence from qualitative studies. We will then run and compare two item response theory (IRT) models; a graded partial credit model (GPCM) and a generalized graded unfolding model (GGUM). We will examine key model fit statistics to determine the best-fitting model and inspect item parameters yielded from model estimations.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Psychometric modeling

128. Character Strengths and Well-being: Does IRT Model Choice Influence the Shape of the Relationship?

Susanna Goosen, Gideon de Bruin
Stellenbosch University

We examined the effect of scoring model on the shape of the relations between character strengths and well-being. In accordance with the "too-much-of-a-good-thing" (TMGT) effect, we argue that the overuse of character strengths can result in negative outcomes, implying a curvilinear relation. Inappropriate scoring models, however, can hamper the detection of curvilinearity. Researchers typically apply dominance scoring models, such as standard item response models, but in many cases, ideal-point or unfolding models are more appropriate. To date, no studies have examined the effect of scoring model on the detection of curvilinearity with respect to character strengths and wellness outcomes. We applied both a dominance (generalized partial credit model, GPCM), and an ideal point model (generalized graded unfolding model, GGUM) to the responses of 1370 volunteer adults who completed the IPIP-VIA (character strengths) and the Mental Health Continuum (well-being). We expected curvilinear relations between character strengths and well-being and we hypothesized that the GGUM would best detect curvilinearity. For each scoring model, we employed hierarchical regression analyses, where, in turn, each character strength was entered in the first step, and the square of the strength in the second step. A reduction in AIC values across steps was taken as evidence of curvilinearity. We also visually compared linear and nonparametric loess regression lines to detect curvilinearity. Using the GPCM, results revealed weak curvilinear relations for six of the 24 character strengths. By comparison, the GGUM detected weak curvilinearity for eight character strengths. For these strengths, the TMGT effect may apply. The results partially support our hypothesis that ideal-point models perform better at detecting curvilinear relations.

WEDNESDAY 3 JULY
Poster Session 2
Topic: Artificial Intelligence in testing, psychological assessment and survey research

199. Gender differences in personality assessment with language indicators from semantic vector subspaces: An invariance study.

José Ángel Martínez-Huertas, Álvaro López-Herranz
National Distance Education University

The measure of psychological constructs, like personality, can be achieved by various approaches through language. Among these, vector space models (VSMs) are one of the most popular tools to formalise them. Within VSMs, we have chosen the Inbuilt Rubric method, which can generate a specific semantic vector subspace for each personality trait. Nonetheless, it is necessary to psychometrically validate its computational scores. In this study, we focused on the Big Five Personality Trait Model. Our aim is to conduct a gender invariance study of the computational scores derived from the five specific semantic vector subspaces, one for each personality trait. These scores are the outcome of the response assessment of 10 open-ended items of a semi structured interview. The participants of this study have been 643 undergraduate students who answered the aforementioned items as, for example: "When I am working with a team, I tend to..." Next, we have fitted unidimensional models from the personality language indicators based on each Big Five trait. Results show that computational scores of Emotional Stability, Openness and Extraversion comparably measure the unidimensional latent factor between women and men, but it is not the case of the means of these variables, displaying metric invariance. Otherwise, this kind of invariance was not found in the traits of Conscientiousness and Agreeableness. In conclusion, we found differences in how these language indicators would capture the personality of women and men. In any case, this is just a preliminary study because we have a limited sample of psychology students with unequal samples sizes in terms of gender, meaning we could have low statistical power and it is not representative of the population. Even so, the results of this study show how gender invariance studies are necessary prior to the analysis of the differences in language assessments between these groups.

WEDNESDAY 3 JULY
Session 4.1 SYMPOSIUM
Topic: Innovations in test development

115. Embedded Standard Setting and Item Alignment in Practice

Ellen Forte
edCount, LLC/USA
Dan Lewis
Creative Measurement Solutions, LLC/USA
Amanda Brice
Curriculum Associates/USA

This symposium will focus on an integrated approach to alignment and standard-setting that enhances coherence among assessment components and offers both time and cost savings for developers. Alignment is the process by which an assessment is created to reflect "what" and "how well" target expectations such that the scores it yields (i.e., scale scores and performance levels) can be interpreted in relation to those target expectations. Alignment evaluation attaches evidence to the question, "To what extent does the assessment yield scores, including performance levels, that can be appropriately interpreted in relation to the target expectations?" (Forte, 2017). This positions standard-setting, which is the process for identifying the points that separate the "how well" levels on the score scale, as a necessary component of alignment quality. Presenters will describe an integrated methodology for establishing and evaluating alignment and setting standards and explain how this approach fits within a Principled Assessment Design (PAD) framework. The first presenter will describe the process for aligning test items to evidence statements articulated in performance levels. These item alignments are best understood as hypotheses at this point in the process. The second presenter will describe the Embedded Standard Setting (ESS) analyses (Lewis & Cook, 2020) that test the item-alignment hypotheses and identify preliminary cut points on a score scale. Presenter three will describe a means of resolution for items with hypothesized alignments that are not supported by empirical data (Lewis & Cook, 2020; Brice, 2021; Lewis, 2024). All three presenters will illustrate how these methodologies and their outcomes and be documented and used in support of a strong validity argument. Audience members will be encouraged to ask questions about the application of the integrated alignment and ESS methods in their own contexts.

Discussant name: Ye
Discussant surname: Tong
Discussant affiliation: National Board of Medical Examiners/USA

WEDNESDAY 3 JULY
Session 4.1 SYMPOSIUM
Topic: Innovations in test development

477. **Embedded Standard Setting: Theory and Practice**

Daniel Lewis
Creative Measurement Solutions LLC/USA
Ellen Forte
edCount, LLC/USA

Embedded Standard Setting (ESS) is not a single activity—it is a set of iterative processes and analyses that occur throughout the assessment development lifecycle under a Principled Assessment Design (PAD) framework. The iterative nature of ESS supports assessment system coherence and provides evidence supporting valid score interpretation and use. ESS advances PAD evidentiary reasoning by requiring the alignment of each item (or score point) to a performance level by the linkage of the item to a specific PLD level and evidence statement (Item-PLD alignment; see Forte, this session). Under ESS, we consider the subject matter expert (SME) alignment of a test item to a performance level as an alignment hypothesis, a proposal subject to verification through analysis of empirical data. ESS formalizes the evaluation of the Item-PLD alignment hypotheses using empirical data to provide three key outcomes. First, cut scores emerge analytically and organically by optimizing the coherence of the Item-PLD alignments and empirical data. Second, impact data—the proportion of students in each performance level—is estimated. The third output from the ESS analysis is a list of ESS-Inconsistent items—items with Item-PLD alignments that are not supported by empirical data. For example, an item aligned to the lowest (highest) level of performance is expected to be relatively easy (difficult). If it is not, then the evidentiary chain of reasoning is broken and the item is subject to review and resolution (see Brice, this session). Thus, under ESS, the evidentiary chain runs not just from the content standards to the test items, but first from the standards to the PLDs, then from the PLDs to the test items and their empirically validated performance level linkages, providing more precise interpretability of the measurement target evidenced by the item. In this Session we will present ESS theory and summarize practical applications of ESS and the associated outcomes.

WEDNESDAY 3 JULY
Session 4.1 SYMPOSIUM
Topic: Innovations in test development

472. Principled Alignment in Support of Validity and Coherence

Ellen Forte
edCount, LLC

Educational assessments generally yield scores meant to be interpreted in relation to a set of academic expectations that encompass both a "what", such as content or skills, and a "how well", which is a degree of mastery or proficiency with the content and skills (Forte, 2023). The "how well" aspect can be conceptualized as degree of sophistication and is often operationalized with both a score on a scale and via performance levels, which may be represented simply as pass/fail, separated by a single point on the score scale, or with three or more distinctive score ranges such as basic, proficient, and advanced. Alignment is the process by which an assessment is designed and developed to reflect both the "what" and "how well" target expectations such that the scores it yields (i.e., scale scores and performance levels) can be interpreted in relation to those target expectations (Forte, 2013, 2017, 2023). This first presenter in the symposium will describe an alignment framework based within a principled assessment design (PAD) approach that clearly addresses "what" and "how well" targets. Using practical examples accumulated from over a dozen alignment workshops with hundreds of panelists, they will describe the steps in the rating process as well as how those rating data are analyzed and used to inform decisions about improving alignment quality. They will further describe how those same ratings can also be used in Embedded Standard Setting (ESS) analyses (Lewis & Cook, 2020) as well as to generate reporting statements to facilitate educators' interpretation and use of the assessment scores. They will conclude by connecting the PAD-based alignment framework to the argument-based approach to validity evaluation and examples of how alignment evidence relates to specific validity questions. The next presenter will address the ESS methodology.

WEDNESDAY 3 JULY
Session 4.2 SYMPOSIUM
Topic: International assessment

156. Raising the Bar: Unveiling the New Quality Standards for PISA

Javier Suárez-Álvarez
University of Massachusetts Amherst
Ava Guez
Organisation for Economic Co-operation and Development

The Programme for International Student Assessment (PISA) measures what 15-year-old students know and can do in reading, mathematics, and science in over 80 countries and economies in 100+ languages. PISA's primary challenge to minimize unjustified variations that threaten comparability lies in enforcing 80+ PISA participating jurisdictions and their contractors to meet PISA Technical Standards. The PISA Technical Standards aim to specify the data collection procedures to create an international quality dataset that allows for valid cross-national inferences. These standards serve as a set of criteria for post hoc data adjudication (decisions on whether the data for a specific country are of sufficient quality for inclusion in the international reports) but do not address other phases of the project cycle (e.g., international coordination, instrument development, analysis, and reporting) or aspects of assessment quality like relevance, validity, fairness, and comparability. In response to this need, the PISA Governing Board commissioned the development of additional standards to support the development of high-quality instruments and the valid interpretation and use of PISA results. The first presentation will describe the rationale and coverage of the new PISA quality standards. The second presentation will describe the principles that underpin the new standards, stressing the similarities and differences with other professional standards. The third presentation will illustrate how the new standards support and strengthen the assessment development of innovative domains in PISA. The discussant will examine to what extent the new Standards reflect the best educational and psychological assessment practices to improve comparability and fairness across cultural contexts. Sufficient time will be allotted for audience Q&A.

Discussant name: Kadriye
Discussant surname: Ercikan
Discussant affiliation: Educational Testing Service

WEDNESDAY 3 JULY
Session 4.2 SYMPOSIUM
Topic: International assessment

504. Purpose and Scope of the New PISA Quality Standards

Ava Guez, Francesco Avvisati, Mario Piacentini
OECD

Over the last two decades, the Organisation for Economic Co-operation and Development (OECD)'s Programme for International Student Assessment (PISA) has become a global yardstick for evaluating and comparing quality, equity, and efficiency in learning outcomes in education systems around the worlds. It has also played a significant role in driving education reforms, allowing policy makers to make more informed decisions. PISA has been at the forefront of educational measurement and has consistently aimed for the highest methodological standards, from assessment design to the reporting of data. However, there has not been, until now, any effort to explicitly articulate these standards for the whole assessment cycle. Indeed, the current PISA technical standards cover the procedures that ensure the consistent implementation of PISA in participating countries and economies, but do not address other phases of the project cycle (e.g. international coordination, instrument development, analysis and reporting) or aspects of assessment quality like validity, reliability, comparability and fairness. Additional standards or guidelines were therefore needed in order to support the development of high-quality instruments and the valid interpretation and use of PISA results. This presentation will provide an overview of the new PISA Quality Standards, including its objectives, coverage, organisation and expected uses from the different target audiences, as well as its development process, focusing on the involvement of PISA participating countries to ensure the production of a document that can effectively strengthen PISA's cross-cultural assessment practices.

WEDNESDAY 3 JULY
Session 4.2 SYMPOSIUM
Topic: International assessment

507. Using the new PISA Quality Standards to strengthen the development of PISA's innovative domain assessment

Mario Piacentini, Ava Guez, Francesco Avvisati
OECD

PISA has included an innovative domain assessment in every cycle since 2012. The innovative domain assessments aim to provide PISA participating countries and economies with a more comprehensive outlook on their students' readiness for life. These assessments drive innovation in PISA by extending its focus beyond the core literacies of reading, mathematics and science and by fostering technological and methodological innovations in the design of the items, analysis and reporting. In prior cycles, PISA has integrated the following innovative domains: Creative Problem Solving in 2012, Collaborative Problem Solving in 2015, Global Competence in 2018, Creative Thinking in 2022. The innovative domain for the 2025 cycle will be Learning in the Digital World (LDW) and will assess computational problem-solving and self-regulated learning. This presentation will follow the life cycle of the LDW assessment as an insightful use-case for the new PISA Quality Standards. Through concrete examples of the development of the LDW assessment, this presentation will show how the specific quality standards provide operational guidelines to support and strengthen PISA's technical quality, indicating expected actions from each actor and providing objective quality-assurance criteria based on procedural and empirical evidence.

WEDNESDAY 3 JULY
Session 4.2 SYMPOSIUM
Topic: International assessment

173. The Role of Fairness, Validity, Comparability, and Reliability in the New PISA Quality Standards

Javier Suárez-Álvarez, Stephen G. Sireci
University of Massachusetts Amherst

The value of the Programme for International Student Assessment (PISA) in informing evidence-based policymaking relies on the degree of precision with which population-level statistics are estimated and reported. But also, the degree to which those aggregate statistics can be meaningfully compared (e.g., country-level mean scores) and the interpretations made based on those comparisons are valid for the intended purposes. Although validity, comparability, and reliability are important components of fairness, they do not address all issues of fairness in assessment. Fairness provides an important additional lens for ensuring the validity of research, policies, and all other aspects of a testing program to promote positive, intended outcomes and minimize negative ones. PISA Technical Standards serve as a set of criteria for post hoc data adjudication (decisions on whether the data for a specific country are of sufficient quality for inclusion in the international reports) but do not address aspects of assessment quality like fairness, validity, reliability, and comparability. This presentation will describe the principles behind the new PISA Quality Standards to deliver relevant, rigorous, and transparent information to policymakers through assessment instruments that provide fair, valid, and reliable data comparable across cultural settings, time, and groups. The presentation will focus on major threats and suggested guidelines for ensuring fairness, validity, comparability, and reliability. The presentation will stress the similarities and differences with other professional standards.

WEDNESDAY 3 JULY
Session 4.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

483. The Rasch/Guttman Scenario Approach to Developing the Scale of Ethical Behaviour in Organisations

Valeriia Manina, Tatjana Kanonire, Alena Kulikova, Maxim Storchevoy
HSE University

The importance of following ethical principles in organizations is growing. Employers educate the organisation's employees about ethical principles and it is important for them to understand whether employees follow these principles and integrate them into their daily routine. However, assessing ethical behaviour is complicated by the high risk of social desirability in responses. The Rasch/Guttman Scenario (RGS) approach to test development reduces the social desirability of responses due to low face validity and provides a more transparent understanding of a construct for respondents. This approach combines the operationalisation technique of facet theory and applies a sentence mapping technique to the item development process. Respondents are given a scenario as a stimulus. In this study, the RGS approach was used to develop a Scale of ethical behaviour in organisations. The Scale consists of 16 scenarios. The scenarios were developed on a matrix of two facets and two levels of extension. In the first step the Scale was tested during cognitive labs. In the second step, data from 255 employees were used for psychometric analyses using the IRT. The results showed appropriate psychometric properties of the Scale. Further development of the Scale and implications will be discussed.

WEDNESDAY 3 JULY
Session 4.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

625. Can we predict Careless Responding? The role of sociodemographic variables

Clara Cuevas-Ureña, Inés Tomás, Ana Hernández, Vicente Gonzalez-Romá
University of Valencia
Anna Brown
University of Kent

Theoretical framework: Online administration of questionaries has become a common practice to collect self-reported data, but with all the advantages (e.g., speed in data collection, larger and more heterogeneous samples), also have come new challenges, such as Careless Responding (CR). CR occurs when respondents fail to give sufficient attention to item content, which leads to poor-quality data (Podsakoff et al., 2012). Previous research has yielded mixed results as to what sociodemographic variables may be influencing CR, being the more common ones that young, poorly educated men tend to respond more inattentive (Álvarez et al., 2019). We intend to shed some light on this question by corroborating the stability of this pattern over time. Objectives: To analyse CR differences depending on gender, age and educational level, and whether the pattern of differences is stable over time. Method: We used a longitudinal design with 3 data collections spaced at 9-month intervals. The initial sample was made of 700 panel respondents (50.4% men, aged between 21 and 59 years, varying from no education up to PhD level). Before testing the effects of sociodemographic variables on CR, we assessed the dimensionality of the CR scale and the potential Differential Item Functioning (DIF) across gender, age and educational level (non-university vs university level). Finally, we modelled the relationship between CR and sociodemographics with Poisson regressions. Results: Results supported the unidimensional structure of CR responses and the absence of DIF. The Poisson regressions indicated that only age and educational level are stable predictors of CR over time. As expected, younger and less educated individuals were more careless. Implications: These preliminary results confirm previous research and open the discussion for differential prevention strategies depending on sociodemographic variables. Study funded by the Spanish Ministry of Science and Innovation PID2022-141339NB-I00

WEDNESDAY 3 JULY
Session 4.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

626. Succeeding in a matrix-reasoning task: effects of cognitive strategy and test characteristics

Natalie Badstuber, Tuulia Ortner
Paris Lodron University of Salzburg

Matrix-reasoning tasks are commonly used in education, hiring, and admissions, and have significant implications for test-takers. Analyzing cognitive strategies provides valuable insights into reasoning processes and task evaluation. Previous research has identified a positive relation between constructive matching and reasoning ability, and a negative relation with response elimination. However, several questions remain unanswered regarding the generalizability of matrix-reasoning tasks and the interplay between cognitive strategies and reasoning ability based on test characteristics. To address these questions, we employed a matrix-reasoning task that has not yet been studied in the context of cognitive strategies. A total of 119 university students (71 women, 76% psychology students) completed 12 items of the Free Response Matrices (FRM) and 12 stem-equivalent items adapted to a single-choice (SC) response format while eye-tracking data was collected. Results revealed limited generalizability in cognitive strategy and performance across different matrix-reasoning tasks. Additionally, we investigated whether eye tracking indicators of strategy could explain variance in the SC score, even when controlling for the FR score. The latency until the first response inspection emerged as the only significant predictor (= -0.18, 95% CI [-0.35, -0.02], p = .028) of SC score. Data supports that response elimination may introduce construct-irrelevant variance into the SC score. The discussion focuses on response format and potential influences due to the test characteristics of the FRM. The implications for test construction, considering the contributions of this study and existing findings, are proposed.

WEDNESDAY 3 JULY
Session 4.5
Topic: Psychometric modeling

114. More equitable and fairer measures of safe environment at schools in Chile

Jorge González
Pontificia Universidad Católica de Chile

Educational research is increasingly incorporating measures of attributes beyond cognitive ability. It is nowadays widely recognized and documented that complementary measures of non-cognitive scores do matter for educational purposes. Although cognitive skills are the most used measures to assess and monitor students' achievement, non-cognitive skills can also be reliable and useful measures to predict academic performance. Despite various studies have addressed the analysis of non-cognitive attributes measured by questionnaires, most of them are not conclusive on the type of scores that should be reported given the multidimensional nature of the latent structure involved. Moreover, almost no study considers the comparability of non-cognitive score measures over time, yet it can be an important need for some educational measurement systems. The aim of this paper is to build a score measure that represents a sub-dimension of a School Climate indicator -the safe environment- and explore different approaches of scale linking to establish the comparability of these measures. We use a sample of 9679 teacher responses in 2017 to 16 polytomous items of the Safe Environment dimension. We explore the use of different multidimensional Graded Response Models (MGRM) to obtain a measure of the safe environment dimension. We also propose two approaches for scale linking when using bifactor-GRM: fixed item parameter calibration, and an extension of the Haebara method to multidimensional models to link the scales of two consecutive years of administration of the safe environment scale. Preliminary results from a simulation study show that the proposed linking methods work satisfactorily. Having comparable measures of safe environment we aim to develop fairer assessment tools that can lead to more equitable learning environments for students.

WEDNESDAY 3 JULY
Session 4.5
Topic: Psychometric modeling

144. A comparison between Rasch Equating Method and Delta Scoring Equating Method using the Saudi National Assessment for Learning Outcomes

Ahmed Haddadi, Mohammed Alqabbaa
Education and Training Evaluation Commission

The Delta Scoring Method (DSM; Dimitrov, 2017,2020), is presently used by the National Center for Assessment (Qiyas) in Saudi Arabia for analyzing, scoring, and equating test forms. The DSM equating procedures are advantageous due to their simplicity, absence of complex computations, and applicability to the Non-Equivalent Groups with Anchor Test (NEAT) design and uses the mean/sigma approach to compute the scaling coefficients (i.e., A and B). This research aimed to contrast the DSM equating method with the Rasch true score equating method, using data from the Saudi National Assessment for Learning Outcomes (NALO) as empirical evidence. To conduct such a comparison, several NALO test forms were subjected to sequential equating using both methods. The results of this investigation suggest a considerable degree of similarity between both equating methods. This conclusion thus validates the application of DSM equating procedures in future practice.

WEDNESDAY 3 JULY
Session 4.5
Topic: Psychometric modeling

448. What can go wrong in a large-scale educational evaluation? Insights and recommendations from an educational assessment in Mexico

Scarlett Escudero,
Facultad de Psicología, Universidad Autónoma de Madrid/Spain
Ramsés Vázquez-Lira, Iwin Leenen
Facultad de Psicología, Universidad Nacional Autónoma de México/Mexico
Miguel A. Sorrel
Facultad de Psicología, Universidad Autónoma de Madrid/Spain

Cognitive diagnostic models (CDMs) are latent class models that have seen an increasing interest in the last decade in fields such as education, clinical and organizational psychology, among others. CDMs are a set of psychometric models that can help determining the attributes or skills mastered by each person, based on their responses on the test, assigning a latent class to each individual. Despite the usefulness of CDMs in many areas, empirical applications are still scarce. Specifically, they have never been applied for large-scale evaluation in Mexico where the assessments had been typically generated and analyzed within the classical test theory and item response theory framework. Therefore, this work presents an innovative empirical application where the CDM framework was used to design, apply, evaluate, and analyze a high school teacher assessment. The assessed dimensions had to do with evaluating teaching knowledge and skills, within the framework of the Nueva Escuela Mexicana. The teaching evaluation test was applied to 8,221 candidates to the national educational system. Although the test and evaluation were developed under the CDM framework from the start and the model was not retrofitted, a common practice in CDMs, the results were not as expected. It was found that the items did not have discrimination and did not perform well. It was also found that almost all examinees were classified in the "extreme" latent classes, that is, in the latent classes that include either none or all measured attributes or skills. The in between latent classes had very little examinees. This possibly relates to the fact that the test was applied during the pandemic and that the contents to be evaluated were not well planned. The purpose of this work is to shed a light on what can go wrong despite having all the necessary elements for a large-scale CDM educational assessment. This work is intended to prevent these issues in future empirical applications with quality data.

WEDNESDAY 3 JULY
Session 4.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

55. Ethical issues in assessment arising from the use of AI

Wayne Camara
LSAC
John Weiner
Life Long Learning
Dragos Illiescu
University of Bucharest
Nancy Tippins
The Nancy T. Tippins Group, LLC

International regulation of AI is under active consideration, but almost no final laws/regulations are in place to guide the assessment industry regarding the ethical and responsible use of AI (ATP, 2022). Exceptions of guidelines pertaining to AI and assessment exist that are either narrowly tailor to a specific application such as employee selection (SIOP, 2023) or a specific organization (Burstein, 2023). The purpose of this proposed symposium is to identify and discuss ethical challenges to the application of AI in assessment and potential solutions. Proposed legislation in the European Union (EU) classifies AI systems by risk and focuses on strengthening rules around data quality, transparency, human oversight and accountability. It also addresses ethical questions and implementation challenges in various sectors ranging from healthcare and education to finance and energy. The level of risk of AI technology would be determined from its potential impact on the health, safety or fundamental rights of a person. The legislation proposes a framework of risk tiers: unacceptable, high, limited and minimal (Feingold, 2023). The Organization for Economic Cooperation and Development produced a framework addressing ethical risks in four areas related to AI: 1. Human rights, privacy, fairness, agency, and dignity 2. Transparency and explainability 3. Security and safety 4. Accountability The proposed symposia will include brief presentations of potential ethical challenges associated with AI in assessment, followed by a panel discussion on how these challenges may uniquely impact international applications of AI in assessment. Audience participation will be encouraged during the panel discussion and Q&A period. The four panelists each have extensive experience in assessment and the development of guidelines. In addition, panelists have written and/or presented on AI issues in assessment.

Discussant name:
Discussant surname:
Discussant affiliation:

WEDNESDAY 3 JULY
Session 4.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

63. Test taker data: Ownership and control of use of data from AI assessments

Wayne Camara
LSAC/USA

AI-based assessments often use test data (responses, keystrokes, latency, routing, tracking, etc.) to develop, refine, or validate AI algorithms and processes. Similarly, AI-based assessments often use test data (beyond results) to make inferences about test takers learning processes, organizational fit, problem-solving strategies, or consistency with different profiles. Even when such uses are in addition to scoring and reporting, they should be explained as part of formal disclosure agreements or statements of uses/purposes. For example, formative assessments may base statements about learning on process data without providing any information to test takers about what data have been used to support such inferences. Laws and standards of professional practice often require formal informed consent for testing and mandate information about test content, purposes, and uses be shared with test takers in advance of testing. Can AI freely use candidate process or secondary-level data for internal improvement or to furnish additional insights without approval when informed consent is required, and without any notice about the additional uses in most assessment scenarios? This discussion will review the unique challenges and responsibilities of AI-based assessments when conveying the use and purposes of assessments even when individual identifying information is protected, as well as statements from the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014) and relevant ITC guidelines.

WEDNESDAY 3 JULY
Session 4.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

64. Security, safety, and accountability in AI-based assessment

John Weiner
Lifelong Learner USA

The use of AI-based assessment amplifies longstanding concerns and raises new ones with respect to data security, test taker safety, and accountability in the adherence to ethical principles (OECD, July 5, 2022). The online delivery of secure exams on test takers' personal devices in their homes, along with the expanding digital footprint for test takers has resulted in the generation and storage of vast amounts of data, including personal information (e.g., identity, geolocation, biometric, and financial data). These data are susceptible to cybercrime and inadvertent data breaches which may potentially result in harm or risk to the test taker (e.g., privacy, financial, physical safety). The recently published Guidelines for Technology-Based Assessment (ITC & ATP, 2022) discuss issues and best practices in operating Test Delivery Environments (Chapter 3) and Data Management (Chapter 6), offering specific recommended processes and procedures delivering secure assessments and managing and protecting test taker data. Accountability is an overarching concern with all facets of ethical AI frameworks and entails having policies in place to determine who is responsible for decisions and outcomes derived with use of AI-based assessment. One of the challenges is that there is no standard accountability mechanism for ensuring that organizations actually follow-up and implement ethical frameworks that they espouse (Hagendorf, 2020). While there are certainly regulations that protect data privacy (e.g., EU GDPR), there is no regulation addressing ethical use in a comprehensive manner; i.e., across the various components of an ethical framework. We will discuss and opine on approaches to assembling best practices and guidelines for ethical AI-based assessment security and safety, drawing from existing sources where available, and suggesting new approaches to operationalizing accountability.

WEDNESDAY 3 JULY
Session 4.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

66. Fairness, transparency & explainability of AI-based assessment

Dragos Iliescu
University of Bucharest Hungary

AI-based assessment has been criticised before for issues related to fairness, transparency and explainability; there are oftentimes confounded to some extent. These issues in AI systems have been well documented, and a number of tools and practices have been developed in time, to help practitioners identify, assess, and control harms caused by AI systems, in one or more of these three categories. Such tools and practices are usually employed as part of AI accountability audits, that are designed and conducted either internally or, for more credibility, by third parties, i.e. external auditors. We will discuss some of these tools, that prompt and target artifacts at different stages of the AI development and usage lifecycle. We will also discuss the extent to which broader knowledge of the biases inherent to AI-based assessment can lead to examples of anticipatory ethics in technology development and usage. More specifically, we will discuss how biases creep into AI systems based on their training data (e.g., training datasets with real-life or synthetic data, or exposure for those systems that keep learning from experience), or training processes (e.g., algorithms used for learning and for predictions, or user interactions). We will then look at various popular fairness metrics (demographic parity, equal opportunity/equal mis-opportunity, average odds) and the way in which they manifest in AI systems; we will specifically discuss the fairness score, and bias index.

WEDNESDAY 3 JULY
Session 4.7
Topic: Computational developments for social science research in cross-cultural testing

243. Advancing Education Assessment through NLP and AI: A Comprehensive Approach to TIMSS and PIRLS

Matthias von Davier
TIMSS and PIRLS International Study Center

This presentation explores the transformative integration of Natural Language Processing (NLP) and Artificial Intelligence (AI) in the context of the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). Focused on three critical components—automated item generation, automated scoring, and automated test assembly—our research delves into the potential of these technologies to enhance the efficiency, accuracy, and overall quality of educational assessments. By leveraging NLP, we aim to demonstrate how AI-driven methodologies contribute to the creation of contextually rich, diverse, and unbiased test items. Furthermore, the presentation addresses the application of AI in automating the scoring process, ensuring swift and consistent evaluation while accommodating intricate linguistic nuances. Finally, we examine the role of AI in optimizing test assembly procedures, streamlining the construction of assessments that align with evolving educational standards. Through this exploration, we aim to highlight the promising prospects of NLP and AI in revolutionizing educational assessment methodologies on a global scale.

WEDNESDAY 3 JULY
Session 4.7
Topic: Computational developments for social science research in cross-cultural testing

379. Titel: Automatic math item generator - Auto.Math: Bridging the gap between tradition and AI?

Steve Bernard, Yannick Rathmacher, Ulrich Keller, Philipp Sonnleitner
University of Luxembourg

More than a decade of research on (model-based) automatic item generation (Gierl et al., 2012; Gierl et al., 2023) has passed and although research has come far, the underlying technology and following implications are still not fully understood, leaving plenty of aspects being under-researched. Meanwhile, items developed by (agnostic) generative AI (LavergheFa Jr, A., & Licato, J., 2023) seem to be the new solution to the time intense and expensive development of test items (Kosh et al., 2018). However, such generated items – despite being cost effective, lack traceability of item components (e.g the stem, the question, distractors, etc.) endangering principles of construct validity. In this presentation, we make the case for not dropping model-based automatic item generation too early by demonstrating and discussing the automatic item generator Auto.Math which is built on psychometrically tested cognitive models. These models were provided by the large, multilingual item pools of the Luxembourg's national school monitoring program (Épreuves Standardisées) which use the national education curriculum as guidance for their item development. Building on the needs for this program and its ever-growing demands for new items the Auto.Math was built. A major feature that distinguishes Auto.Math from others, and especially from AI based models, Is the theoretical framework based on empirically and psychometrically validated data, which allows for the differentiation of certain difficulty levels. This means that all the information entered, the creation process through to the finished item and its attributes is theory-based, transparent, and thus can be traced. We'll discuss fields of application, among them addressing the training needs of pupils, particularly in those areas where national school monitoring programs have found shortcomings. Further testing and validation of the system will be necessary before it can be consider using it for individual assessment purposes.

WEDNESDAY 3 JULY
Session 4.7
Topic: Computational developments for social science research in cross-cultural testing

534. Multiple Imputation of missing values for randomized controlled trials: A step-by-step tutorial using mice

Oscar Lecuona
Universidad Complutense de Madrid
Ariadna Angulo-Brunet
Universitat Oberta de Catalunya
Victor Ciudad
Universitat de València
Ricardo Olmos
Universidad Autónoma de Madrid

Concepual framework: Randomized Controlled Trials (RCTs) are a widely used research protocol in applied research. Among others, a major challenge of RCTs is the presence of missing data due to participant dropout. This leads to loss of power and estimation bias. Multiple imputation (MI) is becoming increasingly popular to deal with missing data in Randomized Controlled Trials. However, MI can produce biased results if not carried out properly. In addition, the required assumptions and steps to develop proper MI for RCTs can be challenging. There is a scarcity of practical guidelines to implement MI for such protocols. Objectives: In this article we provide a step-by-step tutorial on (1) how to assess missing data in a RCT and how to avoid common misconceptions and pitfalls, (2) how to implement MI in and RCT using the mice package, (3) how to analyze RCT data in the MI framework such as implementing linear models, comparison of effect sizes, and plotting results, and (4) how to develop sensitivity analysis to assess robustness and impact of MI. Sample: We illustrate this tutorial with a case RCT for wellbeing in social workers (N = 82) comparing two interventions (mindfulness-based intervention, and wellbeing-based intervention) across four measurements (pre intervention, post intervention, 2-month and 4-month follow-ups). Participants were mostly female (92.7%), single (46.3%) and with undergraduate studies (62.2%). Self-report measurements of depression, anxiety, and mindfulness are used as outcomes. Implications: MI showed stability of the findings and tests used for this dataset, while also the strength of increasing power of conclusions. This tutorial can aid applied researchers to use MI with rigor in their RCT designs. Limitations and extensions of the field are also addressed.

WEDNESDAY 3 JULY
Session 4.8 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

726. **Testing the Periodic Table of Personality Methodology**

Rainer Kurz
HUCAMA Analytics Ltd

The Periodic Table of Personality (PToP) is a ground-breaking innovation that increases understanding of personality construct by mapping scales based on their correlations with orthogonal Big 5 scores based on the Trait Descriptive Adjectives (TDA). This symposium applies the methodology to explore how measurement nuances impact PToP mappings. The first paper uses TDA to map 10 aspects, 40 qualities and 120 personas in Lumina Spark to the PToP. The aspects bifurcate the poles of the Big 5 into separate scales. Each aspect features 4 qualities. Each quality features Underlying, Everyday and Over-extended Persona scales which broadly correspond to inside, outside and dark side measures. The results show how subtle changes to the measurement result in changes in the PToP position. The second paper explores the use of PF16 as an alternative to TDA to create PToP mappings for 48 facets grouped into 8 factors in the broader tool PF48 as well as for NEO IPIP facets and factors. The PToP mappings largely reflect expectations but are less clear cut for 'blended' facets in PF48 and surprising if not disconcerting for some NEO IPIP facets. The third paper explores the impact of ipsatisation on personality assessment. Primary mappings remain intact and in line with expectations. However secondary mappings are all positive for normative scores but are all negative for ipsatised scores. The pattern of secondary loadings backs Cybernetic Big 5 Theory (DeYoung, 2015). The symposium confirms the value and importance of the Periodic Table of Personality and extends the methodological reach and applications.

Discussant name: Richard
Discussant surname: Justenhoven
Discussant affiliation: Welliba

WEDNESDAY 3 JULY
Session 4.8 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

778. Paper 3: Normative and Ipsatised Great 8 Success Factors on the PF16 Periodic Table of Personality

Michele Guarini
HUCAMA Group

This paper explores how positions of Great 8 (Kurz & Bartram, 2002) constructs on the Periodic Table of Personality (PToP) change when normative responses are ipsatised. The impetus for this research is the desire to control 'Response Style' which can make interpretation of normative questionnaires difficult due to overly positive responding (including 'faking good') and overly self-critical responding. This issue is exacerbated in team assessment situations. An item-level ipsatisation method was developed building on Bartram (1996) where items are arranged in 8 virtual octets covering the Great 8. For each normative domain score an ipsatised role was created: INVESTIGATION vs. DEVELOPER; STRUCTURE vs. IMPLEMENTER; SUPPORT vs. ALTRUIST; RESILIENCE vs. OPTIMIST; DRIVE vs. FINISHER; CREATIVITY vs PIONEER; INTERACTION vs. NETWORKER; INFLUENCE vs. INSTRUCTOR; 296 professionals and managers completed PF48 (of which PF16 is a 'core' subset). Five factors extracted from PF16 had Eigenvalues >1 accounting for 68% of the variance. Five PCA Varimax rotated This paper explores how positions of Great 8 (Kurz & Bartram, 2002) constructs on the Periodic Table of Personality (PToP) change when normative responses are ipsatised. The impetus for this research is the desire to control 'Response Style' which can make interpretation of normative questionnaires difficult due to overly positive responding (including 'faking good') and overly self-critical responding. This issue is exacerbated in team assessment situations. An item-level ipsatisation method was developed building on Bartram (1996) where items are arranged in 8 virtual octets covering the Great 8. For each normative domain score an ipsatised role was created: INVESTIGATION vs. DEVELOPER; STRUCTURE vs. IMPLEMENTER; SUPPORT vs. ALTRUIST; RESILIENCE vs. OPTIMIST; DRIVE vs. FINISHER; CREATIVITY vs PIONEER; INTERACTION vs. NETWORKER; INFLUENCE vs. INSTRUCTOR; 296 professionals and managers completed PF48 (of which PF16 is a 'core' subset). Five factors extracted from PF16 had Eigenvalues >1 accounting for 68% of the variance. Five PCA Varimax rotated components were saved for the PToP mappings. Scored normatively, all 8 factor domain scales in PF16 had their highest mapping with the expected Big 5 component with STRUCTURE, RESILIENCE, SUPPORT, and INFLUENCE reaching a factor pure ratio. Secondary loadings were always positive and largely as expected. Using ipsatisation, all 8 factor domain scales in PF16 had their highest mapping with the expected Big 5 component with only OPTIMIST (linked to RESILIENCE) reaching a factor pure ratio. Secondary loadings were always negative with negative loadings on Extraversion for the STABILITY facets, and on STABILITY scales for the PLASTICITY facets. Ipsatisation is known to create a bias towards negative scale correlations (Bartram, 1996). Ipsatised scores fully back the Stability vs. Plasticity differentiation in DeYoung's (2015) Cybernetic Big 5 Theory, and are compatible with Jungian personality theory.

WEDNESDAY 3 JULY
Session 4.8 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

758. Paper 2: Personality Factors on the PF16 Periodic Table of Personality

Rainer Kurz
HUCAMA Analytics Ltd

Woods & Anderson (2016) created the Periodic Table of Personality (PToP) drawing on the Trait Descriptive Adjectives (TDA). This paper explores the use of Personality Factors (PF16) as an alternative. DeYoung, Quilty & Peterson (2007) added Stability and Plasticity meta- factors as well as Aspects that split each Big 5 factor. The associated 10 Aspects scales were used by Guenole (2021) to map the P10 questionnaire to the PToP. This paper uses PF16, an 80-item questionnaire that measures a variation of the Great 8 (Kurz & Bartram, 2002) with two facets each, developed in the light of Cybernetic Big 5 Theory (DeYoung, 2015), to map scales to the PToP. 466 professionals and managers completed PF48 (of which PF16 is a 'core' subset) as part of a large item pool that included 300 items adapted from NEO IPIP. Five factors extracted from PF16 had Eigenvalues >1 accounted for 69% of the variance. Five PCA Varimax rotated components were saved to form the basis for PToP mappings. All 8 factor domain scales in PF48 had their highest mapping with the expected Big 5 component with STRUCTURE reaching a factor pure ratio. 41 of the 48 facets in PF48 had their primary mapping with the hypothesised factor with 9 achieving a factor pure ratio. Empowerment, Self-Esteem, Competitiveness and Trust had their primary and secondary mapping outside their hypothesised component due to their (deliberately) 'blended' nature. All NEO IPIP factors had primary loadings as expected. However, for Agreeableness, Neuroticism and Openness two facets each had their primary loadings outside their hypothesised component. Several 'blended' facets in PF48 had a different primary reflecting deliberate design principles. Outliers on NEO IPIP appear in the light of item content review to be 'rogue' facets that unintentionally veered away from Big 5 orthodoxy. PF16 is an attractive alternative to TDA with 80 rather 100 items, contextual content, and 16 rather than 5 primary scales.

WEDNESDAY 3 JULY
Session 4.8 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

774. Paper 1: Mapping Lumina Spark and Emotion Qualities to the TDA Periodic Table of Personality

Stewart Desson, Jonathan Cannon
Lumina Learning

Introduction This research aims to map Lumina Spark and Emotion to the Periodic Table of Personality (Woods & Anderson, 2016), aiming to provide evidence of convergent validity between Lumina Spark, Emotion and the Trait Descriptive Adjectives (TDA), while also building on the work by Woods and Anderson (2016), by conceptualising opposite sectors in the circumplex model as discrete dimensions. Measures Lumina Spark and Emotion consist of 240 items measuring 40 aligned to the Big Five Factors. The Trait Descriptive Adjectives (Goldberg, 1992) are a set of 100 marker traits for the lexical Big Five by Goldberg (1992). Sample The sample consists of 671 professionals of mixed background. Analyses Principal components analysis with varimax rotation was run on the TDA to identify the orthogonal Big Five factors; regression factor scores were computed as new variables. The Lumina Spark and Emotion Qualities were correlated against the regression factor scores. Primary and Secondary correlations were used as the criteria for mapping, further informed by factor purity and vector length. Results Principal components analysis with varimax rotation of the TDA found that 93 of the 100 adjectives loaded onto their expected factors. Of the 7 that did not load on their expected factor, 3 had their secondary loading on the expected factor. Qualities of Lumina Spark and Emotion were then correlated against the regression factor scores in order to map to the Periodic Table. Primary and secondary correlations were identified, factor purity was assessed, and vector lengths were calculated, with Qualities mostly correlating with their expected factors. Discussion The bifurcated Qualities provided more comprehensive coverage of the Periodic Table, compared to unidimensional conceptualisations, also showing that opposing sectors of the circumplex model could be treated as discrete dimensions. This approach provides greater fidelity to personality assessment compared to traditional models.

WEDNESDAY 3 JULY
Session 4.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

439. Alternative Approaches to Maximising Test Fairness

Jake Smith, Rab MacIver, Sarah Chan
Saville Assessment Ltd, UK

This symposium will present several practical methods that can improve fairness as a means to promote better Diversity, Equity and Inclusion (DE&I) when using psychometric assessments in pre-employment screening. The first paper looks at research concerning the impact of stereotype threat, stereotype priming and the effect this can have on cognitive test performance, with a focus on stereotypes in relation to gender, ethnicity and age. Test performance data is presented to demonstrate the impact of reducing stereotype priming before testing, a simple change to help increase fairness. The next paper considers an approach to reducing the cognitive loading of a situational judgement test (SJT) when being used as a screening assessment. The advantages of employing a single-item rating response format are explored in practice using a very large operational sample. This field study demonstrates how mean score differences between different groups split by ethnicity can not only be mitigated, but completely removed to ensure equal proportion of applicants from different subgroups are progressed through screening. The last paper discusses changing online screening from multiple assessment stages to a single combined stage. The need for an alternative method that organisations can use to robustly model score differences between protected groups is highlighted, as information from organisations' own selection processes is often insufficient to do this reliably with confidence. This paper outlines a methodology for data modelling that can help organisations optimise fairness in screening by combining different assessment scores into a weighted algorithm. This symposium aims to provide practitioners (as well as test publishers) with alternative approaches for implementing cognitive, SJT and behavioural assessment more fairly. Each of the approaches bring an individual gain that collectively can contribute to significantly improved DE&I outcomes in selection processes.

Discussant name: Prof. Hennie
Discussant surname: Kriek
Discussant affiliation: TTS-Top Talent Solutions

WEDNESDAY 3 JULY
Session 4.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

440. **Stereotypes and Testing Fairness**

Sarah Chan, Rab MacIver
Saville Assessment Ltd, UK

Studies on stereotype threat and its impact on test performance in experimental settings (Steele et al., 1995; Shih et al., 1999; Jamieson et al., 2011) find that test-takers from ethnic minority groups, when primed of their ethnicity, can underperform as 'activating' negative societal stereotypes about their group's abilities makes them feel threatened. This raises questions as to whether demographic group differences found in cognitive tests are also partially impacted by stereotype priming. This paper investigates the effect of reducing priming on test performance by moving the position of biodata collection from before sitting an aptitude test, to encouraging candidates to submit their biodata after testing. An operational dataset (N=132,045) of combined aptitude test scores (verbal, numerical and diagrammatic reasoning) were analysed. A series of two-way ANOVA was conducted to investigate the effect of moving the biodata collection position (pre-test, post-test) on the overall test score for different demographic groups split by gender (Male, Female), ethnicity (White, Asian, Black, Other) and age.

WEDNESDAY 3 JULY
Session 4.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

441. Fairer Assessment through Reducing Cognitive Load

Jake Smith
Saville Assessment Ltd, UK

Research literature has consistently observed mean situational judgement test (SJT) score differences between ethnicity subgroups of around 1/3 of a standard deviation (Sackett et al., 2021). Cognitive load – the mental effort placed on working memory – has been identified as a key driver, with increased cognitive load associated with larger differences (McDaniel et al., 2008). While Cohen's standardised mean difference conventions are a useful indicator of the magnitude of differences, in practice, where SJTs are used in selection, relative progression proportions indicate how diversity is maintained in a talent pipeline. The cognitive loading of an SJT has been linked to response scale design. Multiple-choice formats (rank or most/least) typically demonstrate stronger relationships with GMA than rate formats resulting in increased cognitive loading (Glaze et al., 2011; White, 2014; Arthur et al., 2017). This paper considers the use of an SJT for selection by a UK retail company, where a single-item effectiveness rating response format was developed to minimise extraneous cognitive load. Progression rates were calculated for N=352,282 applicants split into subgroups based on ethnicity (BAME, White British, White Other). Results indicated these remained consistent for protected subgroups: • 52% of total applicants were in the 'BAME' group • 54% of progressed applicants were in the 'BAME' group This equates to a relative selection ratio (BAME:White) of 1.08. Given the 4/5ths rule equates to a relative ratio of 0.8, this result provides evidence that completing the SJT has not disadvantaged BAME candidates. In fact, the inclusion of this SJT as part of a selection process has had a positive impact on increasing the proportion of minority applicants progressed to the subsequent stage. These findings are consistent with previous research that SJTs with reduced cognitive load lower (or in this case completely remove) score differences between ethnicity subgroups.

WEDNESDAY 3 JULY
Session 4.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

442. Implementing Fairer Single-stage Online Screening

Rab MacIver, Jake Smith
Saville Assessment Ltd, UK

Online screening of applicants usually comprises multiple stages. Organisations often choose to put shorter, less expensive assessments at the earliest stages. While this may be cost efficient, cognitive tests or situational judgement tests (SJTs) used in isolation as a single stage may result in adverse impact on protected groups (Sackett et al., 2021). By contrast, behavioural questionnaires can be designed to produce scoring algorithms which avoid adverse impact. An established alternative approach is to replace multiple stages with a single screening stage that combines multiple assessments and produces a single score. In practice, when organisations first implement a new online screening process with multiple measures, they often do not have sufficient data from previous processes to be useful in modelling optimal weightings that maximise criterion-related validity and mitigate potential adverse impact. For many well-established processes, there is still insufficient data from minority/protected subgroups (e.g. less than 200 applicants per subgroup) to optimise scoring algorithms for fairness. This paper considers how organisations can implement fair and valid processes without having the requisite usage data. Large samples of existing usage data are used to create simulated models to develop scoring algorithms. Using separate samples of usage data from a cognitive test, a behavioural questionnaire and an SJT, a combined simulated dataset was created to model mean subgroup score differences based on gender, ethnicity and age. The result is the capacity to model different weightings of combined assessment scores and better understand the impact on subgroup progression percentages at different cut scores. This allows organisations to implement optimally-weighted algorithms from different assessments without the need for their own usage data. Methodological limitations and recommendations are discussed.

THURSDAY 4 JULY
Session 5.1 SYMPOSIUM
Topic: Innovations in test development

185. Continuous Norming: Recent Advancements in Research and Application – Part A

Jan-Philipp Freudenstein
Hogrefe Publishing Group

Continuous norming refers to various methods of statistically modeling psychological test scores in relation to predictor variables (e.g., age) to produce standardized test scores. It is an increasingly popular approach for improving the precision of norm scores by avoiding artificial categorization of predictors. In recent years, research has provided several methodological advances in continuous norming that have greatly increased the utility of these procedures for test developers. Despite its success, several challenges and potential improvements remain in the methodology and application of continuous norming. This two-part symposium will address these remaining issues by providing insights into ongoing research and practical applications. The symposium thereby aims to provide a comprehensive understanding of continuous norming as well as in-depth insights into current research and applications. In particular, this first part of the symposium will provide a systematic overview of the current state of continuous norming research. In addition, examples of practical applications of continuous norming will be presented, as well as a proposal for reporting guidelines. Overall, the symposium may guide future research, improve test development practices, and set new standards for reporting in the field of psychological testing.

Discussant name:
Discussant surname:
Discussant affiliation:

THURSDAY 4 JULY
Session 5.1 SYMPOSIUM
Topic: Innovations in test development

188. Zero-Inflated Beta-Binomial Distributions for Regression-Based Norming of Test Data with Floor and Ceiling Effects (Innovations in test development)

Jan-Philipp Freudenstein, Kilian Hasselhorn
Hogrefe Publishing Group

When dealing with floor and ceiling effects in psychological tests, regression-based norming methods often face challenges in accurately estimating norms. This is particularly problematic when test scores are assumed to follow a beta-binomial distribution and test score ranges complicate the use of a normal distribution or its generalizations. We propose the use of zero-inflated beta-binomial distributions for estimating regression-based norms, when easier or more difficult items are not implemented in the test. Using norm data from the IDS-2, we show that, in certain scenarios, regression-based models that follow a zero-inflated beta-binomial distribution provide a better fit to the data and yield improved norm scores compared to a standard beta-binomial model. These findings highlight the importance of considering the distributional nature of raw scores and support the use of zero-inflated beta-binomial models in the norming process, when faced with floor and ceiling effects. Circumstances under which the zero-inflated beta-binomial distribution works well, limitations, and practical guidance for test developers are discussed.

THURSDAY 4 JULY
Session 5.1 SYMPOSIUM
Topic: Innovations in test development

236. Where are we and Where to go? – A systematic Review and Real Data Example of Continuous Norming (Psychometric modeling)

Julian Urban, Vsevolod Scherrer
Trier University
Anja Strobel
Chemnitz University of Technology
Franzis Preckel
Trier University

Norming of psychological tests is decisive for the interpretation of test scores. Conventional norming methods, based on subgroups, can introduce biases, or require very large samples to gather precise norms. New continuous norming methods (inferential, semi-parametric, and parametric norming) propose to solve those issues. We aim to provide insight into the current use of continuous norming methods in test development and tackle the question, which continuous norming method would be preferable. We conducted a systematic review of 121 publications with overall 189 studies and augment this review with an overview of German-language tests utilizing a stratified sample of 52 tests. We further analyzed a real data example (i.e., norming data of a self-report questionnaire assessing Need for Cognition; N = 2,581). We compared the different norming methods in terms of precision and bias and estimated differences between the resulting norm scores. The review revealed that most studies with continuous norms emerged after 2015 and used inferential norming. Continuous norms outperformed conventional norms regarding bias and precision if modelled appropriately. Yet, continuous norming is rarely applied in test development. Moreover, there were several limitations in the appropriateness of modelling and sometimes a poor reporting of the norming process. In our real data example, we found (1) a clear hierarchy of precision in favor of parametric norms, (2) less bias for continuous norms, and (3) large norm score differences depending on the norming method for some individuals. Our study provides insight into current norming practices and adds nuance to prior studies comparing different norming methods. Combining the studies of the review with our empirical findings leads to the conclusion that continuous norming methods outperform conventional methods but the circumstances under which continuous norming methods are preferable are yet unclear.

THURSDAY 4 JULY
Session 5.1 SYMPOSIUM
Topic: Innovations in test development

190. The GRoNC-Checklist: Guidelines for Reporting on Norm-referenced and Criterion-referenced scores (Translation of tests, psychological assessment instruments and survey questionnaire)

Marieke Timmerman
Psychometrics and Statistics, University of Groningen, the Netherlands
Annelies De Bildt
Department of Child and Adolescent Psychiatry, University of Groningen, University Medical Centre Groningen, the Netherlands; Accare Child and Adolescent Psychiatry, Groningen, The Netherlands
Julian Urban
Survey Design & Methodology, GESIS Leibniz Institute for the Social Sciences, Mannheim, Germany; Giftedness Research & Education, Trier University, Germany

Psychological test manuals vary widely in how and what they report upon the construction and interpretation of standardized test scores, on both norm-referenced scores and criterion-referenced scores. This hampers a critical evaluation of the quality and meaning of the standardized test scores involved and, consequently, a correct interpretation of an individual's standardized test score. Because there is no up-to-date, specific standard for how to construct and report upon the construction and interpretation of standardized test scores, we developed Guidelines for Reporting on Norm-referenced and Criterion-referenced scores (GRoNC). We used a systematic approach, based on the approach for the creation of health reporting guidelines, involving 1) a literature search to identify the need for a guideline, and to review the literature on theory on the construction and interpretation of standardized test scores; 2) a preliminary generation of guideline items, based on the literature; and 3) a Delphi procedure among norming experts to reach consensus on the guidelines. We will present the process of developing the GroNC and the GRoNC itself. Additionally, we will discuss its potential to assist test developers to make informed decisions in the process of constructing suitable standardized test scores for their psychological test, encouraging a consistent and a complete reporting of the construction process and a correct interpretation of standardized test scores.

THURSDAY 4 JULY
Session 5.1 SYMPOSIUM
Topic: Innovations in test development

189. Comparing different Continuous Norming models in the creation of Rey Complex Figure Test (RCFT) norms (Psychometric modeling)

Yaiza Puig Navarro
Hogrefe TEA Ediciones

Rigorous standardization of psychological tests and scales constitutes an essential aspect for the accurate interpretation of test scores. Traditional norming methodologies, relying on subgroup data, frequently introduce biases or require very large sample sizes to ensure the precision of normative values. In response to these inherent challenges, continuous norming techniques, such as inferential, semi-parametric, and parametric norming, have emerged as potential solutions. This study focuses on the application of diverse continuous norming methodologies to establish normative data for the sample standardization of the Rey Complex Figure Test (RCFT), a widely utilized neuropsychological assessment tool designed to measure visual memory and visoconstructive deficits in individuals with cognitive impairment. Our research encompassed a sizable cohort of 2503 participants (mean age = 23.76 years; standard deviation = 19.84; 50.3% female). The results highlight the advantages and disadvantages inherent in each continuous norming methodology, underscoring the need to select an adequately tailored approach to the specific sample features. Consequently, the research explores diverse continuous norming modalities in order to select the procedure most aligned to the characteristics of our dataset. In contrast to theoretical approaches, this study emphasizes practical applications so, in addition to the obtained results, factors such as ease of utilization, execution time, and the way results are returned are duly considered.

WEDNESDAY 3 JULY
Session 5.2
Topic: International assessment

376. Understanding higher education students' ethical learning practices

Marcus Henning
University of Auckland

Theoretical/conceptual framework: Academic integrity is crucial to establishing international standards in testing and to optimising learning gains in higher education settings. Three theoretical frameworks are considered to guide this presentation, namely rational choice theory, planned behaviour theory, and situational ethics. Objectives: This presentation critically reviews the results from 11 journal articles that assessed ethical learning practices conducted by the presenter. Sample: Several countries were involved in the research, namely New Zealand (n=366), Iran (n=204), Nigeria (n=330), and Saudi Arabia (n=338). Methodology: The overarching method is thematic analysis of data gained from a series of cross-sectional surveys from multiple global sources. The surveys used numerous methods, such as employing self-report questionnaires, assessing reactions to ethical dilemmas, and requesting qualitative commentaries explaining engagement. Results: The thematic analysis identified five key areas related to the issue of ethical learning practices. These included: (1) rationale for engaging in academic dishonesty, (2) incidence and types of dishonest behaviours, (3) acceptability for engagement, (4) consequences of engagement, and (5) strategies to reduce engagement in unethical behaviours. Implications: Understanding engagement of dishonest behaviours will inevitably lead to the optimal development of strategies to limit their incidence. Educational interventions need to be available to students at the beginning of their learning so that there is no ambiguity regarding what constitutes unethical behaviour. This is required so that students from all cultural backgrounds have awareness of the rules, regulations, and expectations of their educational institutions. There is also a need for more open global discussion to debate this issue to develop common ground regarding what is permissible learning practice to guide international standards in testing.

WEDNESDAY 3 JULY
Session 5.2
Topic: International assessment

510. Adopting ITC Guidelines for cross-cultural assessment: Translation and validation of an ADHD measure

Braden Hansma, Janine Victor, Joanna Solomon
MHS

Background: Translating and adapting psychological assessments for international use can be a challenge. Luckily, the Guidelines for Translating and Adapting Tests (International Test Commission, 2017) provide an excellent framework for accomplishing such a feat. As a test publisher interested in bringing our assessments to a global audience, we have adopted the guidelines for all international translations. The following study covers our experience applying these guidelines to the development and validation of translated versions of a prominent adult Attention-Deficit/Hyperactivity Disorder (ADHD) assessment, the Conners Adult ADHD Rating Scales 2nd Edition (CAARS™ 2; 2023). Method: Cultural and linguistic translations were created for France, Spain, and Japan. Further, the applicability of the original North American English version was examined in the United Kingdom, Australia, and South Africa, and the European Spanish version was examined in Colombia. Data were collected for Self-Report and Observer forms in the 7 different countries in 2023 (N ranged from 354 to 373 for each rater form for each country). Samples were matched on age and gender to the North American Normative Sample, and the psychometric properties of the different versions were compared for equivalence. Results: Analyses, including differential item/test functioning, omega reliability estimates, and mean scale score differences, will be discussed, addressing similarities and differences between each of the regions as compared to the matched North American samples. Conclusion: The guidelines created a structured and comprehensive framework for developing and adapting international translations. In our experience, adoption of the guidelines led to the successful creation of several international versions of the CAARS 2. Reflections, implications, and interpretations of the results will be discussed.

WEDNESDAY 3 JULY
Session 5.2
Topic: International assessment

567. The Instructional Sensitivity of Constructed-Response Items in International Large-Scale Assessments

Anne Traynor
Purdue University/United States
Özge Altıntaş
Ankara University/Türkiye
Yu-Hui Chang
National Sun Yat-sen University/Taiwan
Setlhomo Koloi-Keaikitse
University of Botswana/Botswana

Understanding and improving student performance in international large-scale assessments requires a nuanced exploration of the factors influencing test outcomes. The instructional sensitivity of the items may be a crucial dimension, shaping the link between teachers' instruction and individual item-level test performance. The term "instructional sensitivity" refers to the degree to which students' assessment item responses reflect the impact of classroom instruction. This study aims to characterize the content features of items with high and low instructional sensitivity, which have been identified previously using a statistical index. Expert raters, comprising master science teachers from Botswana, Taiwan, Türkiye, and the United States, will contribute judgmental item content analysis data to uncover patterns in instructional sensitivity. This study aims to explore how the content of teachers' instruction may directly influence students' performance on individual test items. Hence, Grade 8 or Form 2 science teachers from each nation will be recruited, with at least 5 years of upper primary/lower secondary science teaching experience and the highest educational credentials for their respective grade levels. Participants will engage in structured interviews, predicting students' cognitive response processes for 24 test items and comparing content features of 8 items with high (or low) instructional sensitivity. The study items are owned by the International Association for the Evaluation of Educational Achievement (IEA) and released to our research team for analysis. We will conduct reflexive thematic analysis of recordings from teachers' item content analysis interviews, and intend to draw conclusions about science test item design. Through this exploration, we anticipate contributing valuable insights about assessment design and the evaluation of international educational systems.

WEDNESDAY 3 JULY
Session 5.2
Topic: International assessment

613. Do students respond inconsistently on mixed-worded scales in the PISA 2022 questionnaire?

Michalis Michaelides, Evi Konstantinidou
University of Cyprus

Mixed-worded scales are widely used in assessment instruments to measure constructs of interest. Positively and negatively keyed items within a scale encourage participants be more thoughtful when responding. However, evidence suggests that reversed keying generates wording effects and reduces score reliability and validity. At the level of the individual, those who fail to switch their answers in reversed items produce inconsistent responses. The proposed study aims to investigate inconsistent responding in the 2022 Programme for International Student Assessment (PISA) across 6 country samples. The prevalence of inconsistent respondents will be examined on two mixed-worded scales: the 6-item Sense of Belonging and the 10-item Stress Resistance Scales. Both scales were balanced with equal numbers of positively- and negatively-valenced items, the former appearing early, and the latter appearing later on the PISA student questionnaire. Data from a diverse group of countries in terms of mean performance, geographic location, cultural and socioeconomic level were used: Brazil, Denmark, Greece, Saudi Arabia, Singapore, and the United States. Confirmatory factor analysis showed that a unidimensional factor structure could not be supported for the two scales. Addition of an orthogonal, secondary latent factor which loaded on the negatively-valenced items improved the fit, suggesting a keying effect may result in differential response tendencies. Further analysis will attempt to fit a Factor Mixture Analysis model on each sample to classify examinees into consistent and inconsistent response behavior. Results will provide an insight into (a) the extent of inconsistent responding and (b) how this generalizes cross-culturally. In addition, we will examine (c) if inconsistent behavior is greater towards the end of the questionnaire, and (d) whether the classification into the inconsistent category is stable across two scales differentially positioned within a long survey.

WEDNESDAY 3 JULY
Session 5.2
Topic: International assessment

797. Integrating Intelligent Tutoring Systems with Active Learning to Enhance Testing Systems

Yoon Soo Park
University of Illinois College of Medicine

Theoretical Framework: Intelligent tutoring systems (ITS) provide the promise of self-paced, personalized, and adaptive feedback to students based on individualized performance. ITS incorporates concepts in artificial intelligence and cognitive learning theory to generate powerful learning tools of feedback through educational technology. The integrated testing system uses pre-class ITS tasks with in-class active learning, through a flipped-classroom approach. Objectives: This study examines the use of the integrated intelligent tutoring systems with active learning, administered in Vietnam and in Uruguay across students in mathematics (elementary and middle school) using an international testing system. Sample: Data from 4,072 students were collected across two data collection waves, prior to use of the integrated ITS and post implementation across 6-month intervals in multiple schools across Vietnam (7 schools) and Uruguay (108 schools). Methodology: Data were gathered using multiple data sources, including the adaptive ITS and large-scale mathematics assessment as outcomes. Mathematics performance was estimated using item response theory. Mixed effects regression models were used to examine changes in student performance, including the testing of interaction effects between ITS performance and mathematics scores. Results: Findings show that the combined ITS and active learning show robust effects in student outcomes, with the interaction significantly improving performance by 0.40 standardized units. Increased episodes of ITS completion were significantly related to performance. Results also showed improvements in student and teacher efficacy. Implications: Study results demonstrate combining intelligent tutoring systems with principles of instruction and active learning to generate more robust and enhanced student outcomes. The integrated paradigm proposed may be used to generate greater impact for large-scale international assessments.

WEDNESDAY 3 JULY
Session 5.3 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

425. Comparability challenges in national and international summative assessment

Alejandro Veas
University of Murcia/Spain
Louise Badham
International Baccalaureate Education Program/United Kingdom
Heather Kayton
University of Oxford/United Kingdom
Elena Govorova
Estudios y Evaluaciones 2e/Spain
Jennifer Pérez-Sánchez
University of Salamanca/Spain

Comparability is a central concern in large-scale summative assessments. Final scores have a significant impact on students' future academic and professional opportunities, and are used by teachers, school leaders and policymakers to implement change in educational policy and practice. It is therefore essential that rigorous processes are in place to investigate comparability in large-scale assessments both within and across subjects, so that public trust in assessment systems can be maintained. Comparability is consequently interconnected with validity, as evidence of comparable standards must be gathered so that valid inferences can be made about final scores. However, this is a complex and multifaceted area of assessment that comes in many guises, including the comparability of assessments over time, across regions, in different language versions and different subject areas. This symposium offers a curated collection of presentations from national and international assessment contexts, united by the common thread of exploring comparability issues. The symposium comprises four presentations and a discussant. The first presentation discusses inter-subject comparability in Spanish university entrance exams (EBAU), in particular on variations between different Spanish autonomous communities. The second presentation goes on to discuss the predictive values of PISA assessment results for Spanish university entrance exams. The third presentation discusses inter-subject comparability challenges in the International Baccalaureate, particularly in the case of small cohort subjects and those with complex linguistic profiles. The fourth presentation examines comparability across different language versions of PIRLS assessments in South Africa.

WEDNESDAY 3 JULY
Session 5.3 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

449. Application of the Rasch model to the Spanish university entrance examinations (Testing equivalence by psychometrics methods)

Elena Govorova
2E Estudios & Evaluaciones

Numerous studies have highlighted the need to analyse the effectiveness of academic certification tests designed for university entrance, and to guarantee the comparability of the results. This research analyses the quality and comparability of the Spanish University Entrance Examinations (PAU - Pruebas de Acceso a la Universidad) using the theoretical framework of the construct comparability approach. This approach assumes that it is possible to compare the qualifications obtained by students in the different subjects involved in the university selection process. The main objective is to assess the comparability of the PAU exams across the autonomous communities of Spain, focusing on the relative difficulty of these exams to determine the feasibility of cross-regional comparisons of results. The final scores of over 900,000 students in core subjects such as Spanish Language and Literature, Foreign Language (English), Mathematics and Spanish History, as well as 100,000-300,000 students in ten additional subjects such as Co-official Language (e.g. Catalan, Galician, etc.), Biology or Latin, were analysed over a five-year period (2017-2021). Using the IRT Rasch model, the pooled data for the observed variables (considering the grades for each subject as items of an academic achievement instrument) in each academic year and autonomous community were independently calibrated. Item difficulty parameters and model infit/outfit indices were computed and compared. The results show significant differences in the relative difficulty of the entrance examinations among the Spanish autonomous communities. Furthermore, intra-regional variations in item difficulty were observed, demonstrating inconsistency across regions. The study highlights the need to re-evaluate the formulation and validation processes of university entrance examinations in Spain. This reassessment is crucial to improve the comparability of university entrance exams.

WEDNESDAY 3 JULY
Session 5.3 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

479. Evaluating the cross-language comparability of PIRLS in South Africa (Testing equivalence by psychometrics methods)

Heather Kayton
University of Oxford

PIRLS 2016 was conducted in 11 language versions in South Africa. Overall reading performance differed substantially across language groups, with students tested in African languages severely underperforming relative to English and Afrikaans groups. Moreover, average achievement across all languages was more than a standard deviation below the international PIRLS average, threatening the reliability and validity of inferences made from the results. Given the potential policy implications of results from large-scale assessments such as PIRLS, it is essential to rigorously evaluate whether the assessment provides truly comparable measures across linguistically diverse populations. This paper investigates the comparability of three language versions of PIRLS Literacy 2016 in South Africa. Differential response functioning (DRF) techniques are applied to evaluate comparability for the English, Afrikaans, and Sepedi language groups. Significant DIF was found for 28% of items between Sepedi and English, and 20% between Afrikaans and English. Once the presence, magnitude and direction of DIF was established at an item level, the overall impact of DIF on score comparability was evaluated at both the passage and test level. The impact of DIF at a passage level revealed that 11 passages functioned differently between Sepedi and English groups, and 4 passages functioned differently between Afrikaans and English. At the test level, the DIF did not result in significant test-level differences between the Afrikaans and English versions, however, significant overall divergence between Sepedi and English versions emerged. Furthermore, substantial concerns were found regarding the targeting of item difficulty to student ability in the South African sample. These findings underscore the need to rigorously evaluate cross-language comparability for large-scale assessments, especially in contexts where substantial educational inequality exists.

WEDNESDAY 3 JULY
Session 5.3 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

426. University entrance examinations and PISA: Analysis of curriculum and competences from the Spanish context (International assessment)

Jennifer Pérez-Sánchez,
University of Salamanca
Alejandro Veas
University of Murcia

Educational assessments serve as a valuable resource for enhancing selection processes for university entrance, future employment and other purposes. The Programme for International Student Assessment (PISA) globally evaluates students' ability to use their reading, mathematics and science knowledge and skills to address real-world challenges. It is a valuable tool of academic achievement and educational quality. PISA outcomes have demostrated their predictive capacity for promotion decision and academic grades. In Spain, the current university entrance examinations are referred to as EBAU (Evaluación del Bachillerato para el Acceso a la Universidad). This study aims to examine comparatively the academic achievement of Spanish students using data from PISA (2015 and 2018) and EBAU assessments (2017 and 2020). Data analysis was carried out using linear regression analyses. We found a significant relationship (p

WEDNESDAY 3 JULY
Session 5.3 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

447. Inter-subject comparability: small entry subjects and minority languages in the International Baccalaureate Diploma Programme (International assessment)

Louise Badham
International Baccalaureate
Mkululi Wami
University of Oxford
Antony Furlong
International Baccalaureate

Comparability is a core assessment value for the International Baccalaureate (IB), that is fundamentally entangled with issues of fairness and validity. Students' final grades in the IB Diploma Programme are expected to represent equivalent levels of attainment so valid inferences can be made for university recognition and other purposes. The IB therefore produces statistical estimates of the relative difficulty of different subjects, by comparing mean grades across common candidates in pairs of subjects. This evidence is used to inform standard setting, with the aim of ensuring inter-subject comparability. However, the current approach requires sufficient data to generate reliable estimates of subject difficulty. Yet, more than 150 IB subject offerings do not meet the data requirements as they have fewer than 100 candidates. Moreover, most of these subjects comprise minority language gap in current processes. The current study explores alternative approaches to investigate comparability in these small-entry subjects and minority languages. This paper presents preliminary findings from a mixed methods research study. Qualitative findings from interviews with assessment staff unpack specific comparability challenges in small-entry subjects and minority languages as well as the practices that have been developed to address them, such as cross-language standardization meetings. Furthermore, using a multilevel modelling approach, a secondary data analysis of historical IB assessment data will explore the impact of factors such as the language of instruction on overall DP point scores within the complex linguistic contexts in small-entry subjects. The presentation concludes with recommendations for qualitative and quantitative approaches for investigating inter-subject comparability in small-entry subjects, and the linguistic minorities represented within them.

WEDNESDAY 3 JULY
Session 5.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

240. Test translation and adaptation – survey methodology meets testing

Dorothée Behr
GESIS - Leibniz Institute for the Social Sciences

SUMMARY: Collecting reliable, valid, and, on top of that, comparable data is the requirement for drawing sound conclusions in cross-cultural research. Over many years, scholars across various disciplines have tackled and refined translation procedures and pre-conditions, learning from past experiences, other disciplines, or experiments. Attention is no longer restricted to the proper translation and adaptation step but has also reached the stage of source development, the latter to pave the way for translatable and culturally relevant instruments in a cross-cultural setting. In this symposium, four dimensions of the production of sound translations of measurement instruments are presented: (1) Advance translations conducted on pre-final source instruments to evaluate and improve the translatability of instruments; (b) translation guidelines, notably item-specific ones, to clarify the measurement goal of items and foster comparability across the different language versions of an instrument; (c)) translation procedures and their evolution, and (d) the background of translators involved in translation and adaptation activities. The goal with these presentations focusing on important developments within cross-cultural survey methodology is to foster and encourage cross-fertilization across disciplines.

WEDNESDAY 3 JULY
Session 5.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

368. Best practice in translation of survey instruments: 25 years of practice and research (Translation of tests, psychological assessment instruments and survey questionnaire)

Alisú Schoua-Glusberg
Research Support Services Inc.

Best practice in translation of survey instruments: 25 years of practice and research In the last quarter of the 20th century, survey research began to look into alternative ways to translate and assess questionnaire translations. Back translation appeared as a promising approach and became best practice for a couple of decades. Toward the end of the 1990s, dissatisfaction with this approach led survey researchers with linguistics background to consider and experiment with alternative approaches. This presentation will cover the history of survey translation, problems with back translation, early team translation approaches (e.g. Brislin, Modified Committee Approach), Harkness' (2002) TRAPD model, and implementations of TRAPD, currently considered best practice in the translation of data collection instruments. We will include a discussion of recent literature with experiments comparing different methods and approaches.

WEDNESDAY 3 JULY
Session 5.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

325. Advance translation - a method to enhance source questionnaire's translatability and cross-cultural portability (Translation of tests, psychological assessment instruments and survey questionnaire)

Brita Dorer
GESIS-Leibniz Institute for the Social Sciences

To achieve high-quality translations, also the source text, that is, the text out of which is translated, matters. The same applies when translating questionnaires. The method of Advance Translation has been developed to minimize such problems: a pre-final version of the source questionnaire is translated with the purpose of pointing out later translation problems. Translation is thus used as a problem-spotting tool to enhance a source questionnaire's translatability and cross-cultural portability. Following the "committee" or "team approach", experienced questionnaire translation teams, consisting of questionnaire translators and survey experts, translate pre-final source items and comment on any problems they encounter when transferring the source text into their target languages and cultures. These findings are then incorporated when finalising the source questionnaire, before it will be translated into all target languages. The method has been successfully applied for several rounds in some large multilingual survey projects. Its usefulness was confirmed in a think-aloud study for a specific use case. The presentation will first explain the method and its background and how it has been implemented in different multilingual survey projects. Then we will present examples of source text elements changed because of Advance Translation comments, as well as results of the think-aloud study for assessing the method's usefulness. In the discussion, we will look at similar methods and compare each method's pros and cons.

WEDNESDAY 3 JULY
Session 5.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

262. Going beyond translation procedures – supporting comparability using (item-specific) translation guidance (Translation of tests, psychological assessment instruments and survey questionnaire)

Dorothée Behr
GESIS - Leibniz Institute for the Social Sciences

In 2023, Clifton and colleagues, from within the field of psychology, called for increased efforts to provide "scale-specific information" to translation teams to compensate for what other more generic translation guidelines cannot offer. Such guidance should typically be produced during source scale development, and regularly amended with feedback from translation experience with a given scale. Called differently (e.g., translation notes, annotations or guidelines), similar efforts have been implemented in large-scale cross-cultural surveys for many years (Behr & Scholz, 2011; Harkness, Pennell, & Schoua-Glusberg, 2004). In this presentation, we will provide an overview, spanning across disciplines and studies, about the types of guidance and background information that can (and should) be provided to translation teams, notably focusing on item-specific translation notes/information. Strengths and weaknesses are discussed, as well as use cases presented.

WEDNESDAY 3 JULY
Session 5.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

401. Questionnaire translation in cross-cultural research: Translators, their background, and relevant competencies from the perspective of international research teams (Translation of tests, psychological assessment instruments and survey questionnaire)

Ulrike Efu Nkong
GESIS Leibniz Institute for the Social Sciences

Measurement instruments in different languages are at the heart of cross-cultural survey research. To ensure comparability and comprehensibility of these different language questionnaires and thus enhance data quality, much research has been done within the past years on best-practice recommendations for translation procedures. However, the role and qualifications of the personnel involved in producing these questionnaire translations has not yet been examined in detail. This study aimed at surveying the current status quo of questionnaire translation in cross-cultural survey programs and their varying methodologies. Early in 2023, we used a web survey to collect information from participating country teams of 13 large-scale international survey programs to find out more about their translation procedures and personnel. The initial translations of questionnaires (prior to any reviews or checks) are often produced exclusively by members of the research team or other researchers (44%) or exclusively by professional translators (29%). Less teams (19%) combine the expertise of translators having different backgrounds. About one third of the country teams produce only one initial translation per language. Thus, international best practice recommendations are not necessarily taken into consideration. There seems to be a correlation with an available translation budget, but other reasons, like the importance of subject matter and questionnaire design knowledge vs. the importance of translation expertise (as perceived by the country teams) also seem to play an important role in translator selection. Also, significant impacts of certain levels of experience on the satisfaction with translators' performance could be observed. This presentation will allow an insight into the results and analysis of this study. Further research is currently being conducted on the actual impact a translators' background may have on the translation product and the resulting data.

WEDNESDAY 3 JULY
Session 5.5 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

742. **Assessment in Organizational contexts**

María Dolores Nieto Cañaveras, José Muñiz
Nebrija University

Nowadays, evaluation in the organizational environment involves the simultaneous assessment of several variables that have progressively come to the fore to become interrelated protagonists. Taking this in consideration, this symposium aims to present and illustrate the study of the psychometric properties of a set of reliable and valid tools to measure three of the most relevant variables in organizational settings in the current scenario: work engagement, organizational climate, and entrepreneurial personality. Specifically, the first presentation aims to provide an extra-short measure of work engagement (i.e., Mini-ESCOLA) to be used in empirical and research contexts. The second presentation illustrates the entire process of analysis of the internal structure of a new measure of organizational climate using novel and sophisticated psychometric procedures. In turn, the third presentation relates the above-mentioned measures to study, a) whether work engagement is an adequate predictor of organizational climate, and b) analyze the possibility that both constructs combined allow an accurate prediction of wellbeing at work. Finally, the fourth presentation provides a battery (i.e., BEPE) to assess eight traits related to entrepreneurial personality. It is worth mentioning that findings in the first three presentations are linked via the same data set containing responses from more than 500 workers from the University context (teaching and non-teaching staff). In turn, the fourth presentation is based on a huge sample of workers, leading to sound results. Overall, we are confident that these tools will be of interest when assessing the mentioned variables in organizational contexts.

Discussant name: Ana
Discussant surname: Hernández-Baeza
Discussant affiliation: University of Valencia

WEDNESDAY 3 JULY
Session 5.5 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

744. Work Engagement revisited: A three item express scale (Translation of tests, psychological assessment instruments and survey questionnaire)

María Dolores Nieto-Cañaveras, María Dolores Notario-Jurado, Luís Díaz-Marcos, José Muñiz
Nebrija University

Work engagement is one of the factors that has gained most attention in the last decade in organizational settings. This is mainly due to the impact it has been found to have on work performance. In such context, it is common to measure and examine the connection between several variables related to the organization and the workers themselves. Therefore, the main purpose of this study is to develop a brief version of the Work Engagement Scale (ESCOLA) to provide reliable and valid measure of work engagement to be used in practical and research contexts. Responses from more than 500 workers from the University context (teaching and non-teaching staff) were collected and used to advance the knowledge regarding the latent structure of the ESCOLA scores. The set of analyses conducted included the assessments of redundant items (using unique variable analysis), dimensionality (with parallel analysis and exploratory graph analysis), factor structure (via exploratory structural equation modeling), measurement invariance (e.g., across gender, age, and job position), and reliability. Finally, a three-item Mini-ESCOLA version was obtained, which showed a generally good performance. Due to its good psychometric properties and short length, the Mini-ESCOLA constitutes a valuable tool to be administered in organizational settings where an efficient and accurate measurement of work engagement is needed.

WEDNESDAY 3 JULY
Session 5.5 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

745. Dimensionality of Organizational Climate (Translation of tests, psychological assessment instruments and survey questionnaire)

María Dolores Nieto-Cañaveras, María Dolores Notario-Jurado, Luís Díaz-Marcos, José Muñiz
Nebrija University

Organizational climate has become a core variable in work settings, with prior studies connecting it with other fundamental constructs in this area such as job satisfaction, psychological well-being, and more recently work engagement. Traditionally, the internal structure of organizational climate has been subject of controversy. This has been reflected in the existence of various instruments to evaluate this construct showing heterogeneous structures. With this in mind, the aim of this study was to examine the factor structure of a new scale to measure organizational climate in university education settings using a novel network psychometrics approach in combination with an exploratory structural equation modeling framework (ESEM). The sample from the previous study was used. After using unique variable analysis to identify and remove redundant items that could distort dimension recovery, the analyses comprised the assessments of dimensionality (combining exploratory graph analysis and parallel analysis and), factor structure (using exploratory structural equation modeling), measurement invariance (e.g., across gender, age, and job position), and reliability. Finally, main findings and practical implications of the study are discussed.

WEDNESDAY 3 JULY
Session 5.5 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

750. Wellbeing at work can be measured (Translation of tests, psychological assessment instruments and survey questionnaire)

José Muñiz, María Dolores Nieto-Cañaveras, María Dolores Notario-Jurado, Luís Díaz-Marcos
Nebrija University

Interest in evaluating well-being of workers is growing in the last years. Concurrently, work engagement has been gaining relevance and has proven to be related to other factors that affect well-being in the work context such as organizational climate. In view of this scenario, the objective of this study is to analyze the relationship between perceived organizational climate (understood as different variables of the work context) and work engagement in university workers while considering the potential moderating effect of individual factors (gender, age, seniority, job position). Additionally, we will investigate the possibility that both constructs combined (organizational climate and work engagement) allow an accurate prediction of wellbeing at work. Responses from the sample and instruments used in the studies above (the new organizational climate scale and the Mini-ESCOLA) were used. A structural equation modeling approach was used for the analyses. In line with prior studies, it was found that work climate scores have a positive impact on work engagement, supporting the fact that the measures used can be used together as a measure of well-being at work. Main findings and practical implications of the study will be finally discussed.

WEDNESDAY 3 JULY
Session 5.5 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

754. Entrepreneurial Personality Assessment: The BEPE Battery (Translation of tests, psychological assessment instruments and survey questionnaire)

Álvaro Postigo, Marcelino Cuesta
University of Oviedo
Javier Suárez-Álvarez
University of Massachusetts Amherst
Luis Manuel Lozano
University of Granada
Eduardo García-Cueto
University of Oviedo
José Muñiz
Nebrija University

Entrepreneurial behavior has great personal, social and economic relevance in organizational environments. This behavior depends on several factors, one of which is the entrepreneurial personality. Previous research distinguishes two main approaches: those who choose to assess entrepreneurial personality through the general traits, such as the Big Five personality test, and those who use more specific traits. The present study aligns with the latter approach, and the main objective is to present and describe the psychometric characteristics of the Battery for the Assessment of Entrepreneurial Personality (BEPE). The BEPE evaluates eight traits related to entrepreneurial personality: self-efficacy, autonomy, innovation, internal locus of control, achievement motivation, optimism, stress tolerance, and risk-taking. BEPE shows adequate psychometric properties in terms of reliability and evidence of validity. Additionally, the short and computerized versions developed have also shown suitable psychometric properties, thus serving as complementary versions to the original BEPE. The BEPE can be a useful tool in the organizational context to assess personal aspects of candidates in personnel selection and workers in intra-entrepreneurship matters, thereby aiding decision-making within the organization.

WEDNESDAY 3 JULY
Session 5.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

101. Comparing Content Validity in Personality Assessment: Machine vs. Human-Authored Items using LLMs.

Simon Baron, Emeric Kubiak, Tales Marra
AssessFirst

Objective. As the field evolves, machine learning and NLP techniques are introducing the potential for automatic item generation. However, it raises questions about the efficacy of algorithm-driven methods compared to traditional human-driven processes. Our study aims to compare the content validity of machine and human-authored personality items measuring HEXACO traits. Method. In study 1, we utilized a range of prompt-based generative pre-trained transformers, including GPT-3.5, GPT-4, GPT-3.5-turbo, GPT-4-1106-preview, GPT-3.5-turbo-1106, Llama2:13b, Openchat:7b-v3.5, Zephyr:7b-beta, and Mistral:7b. Each model was tasked with generating 20 personality items for each of the six traits. A panel of 3 psychologists was given an identical assignment. We evaluated the content validity of these items using transformer models, as outlined in Fyffe et al. (2023). In Study 2, we focused on the best performing LLM and on items that were misclassified, challenging the LLM to enhance and refine them. Results. In Study 1, human-authored items consistently outperformed machine-authored items across all metrics, including average accuracy (0.92), precision (0.91), recall (0.92), and F1-Score (0.91). Among the machine-authored items, Zephyr:7b-beta achieved the best performance, surpassing Mistral-7b and GPT-4. Llama-13b showed the lowest performance in all metrics. Notably, both human-authored and machine-authored items yielded the lowest results for the Humility trait. In Study 2, the performance of Zephyr:7b-beta improved marginally, showing a 3% increase in accuracy, precision, and F1-score, and a 2% in recall. Conclusion. Our findings demonstrate that while machine learning show promise in generating personality assessment items, human-authored items still outperform in terms of content validity. This underscores the importance of human expertise in the development of personality assessments, especially for nuanced traits like Humility.

WEDNESDAY 3 JULY
Session 5.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

415. Using ChatGPT for Semi-Automatic Generation of Items for Summative Assessment: a Case Study

Sergio Araneda, Christopher Foster, Susan Weaver
Caveon Test Security

In this study, we explore the use of Large Language Models (LLMs) like GPT-3.5 in the automatic item generation (AIG) for testing, particularly in summative assessments. The project aimed to create about 1,600 items for a client's test, starting with initial drafts by GPT-3.5 and further refined by human review. The paper provides insights into the workflow, methodology, challenges, and preliminary insights into the efficiency and implications of using AI for item development. The project involved developing an application integrated with the Scorpion™ platform, allowing test developers to create and review items using GPT-3.5. It also covered the project's different stages, including item creation, review, and editing, involving content like slide decks, text, videos, and activities from various courses. The methodology included both qualitative and quantitative analyses, with the latter showing an overall 66% acceptance rate of AI-generated items by subject matter experts (SMEs). Significant findings include the necessity of high-quality, deep content for successful AI-generated items, the tendency of AI-generated items to be somewhat monotonous, and the evolution of the role of SMEs from creators to reviewers. The paper concludes that AI can significantly speed up the item construction process and suggests a potential future where AI-assisted item development becomes more prevalent, leading to reduced costs and larger item pools. Despite the efficiencies, the study emphasizes the need to keep human experts in the loop to ensure quality and reliability.

WEDNESDAY 3 JULY
Session 5.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

503. Unlocking Cognitive Patterns: A Comparative Exploration of Linear and Deep Dimensionality Reduction Approaches in clusterizing Student Cognitive Profiles

Matteo Orsoni, Sara Giovagnoli, Sara Garofalo, Noemi Mazzoni, Matilde Spinoso, Mariagrazia Benassi
University of Bologna

Cognitive profiling plays a crucial role in understanding learning dynamics, it contributes significantly to the development of students' metacognitive skills and awareness of the learning process, thereby facilitating the adoption of tailored learning experiences. Clustering, proves effective in cognitive profiling. However, the challenge of the "curse of dimensionality" introduces complexities that can impact the accuracy of cluster subject attribution. This paper investigates the evaluation of various cluster internal validation metrics and cluster stability using a dataset of 1626 participants comprising 54 items across six cognitive domains from the digital assessment tool, PROFFILO. We employ three clustering procedures—K-means, Gaussian Mixture Models, and Fuzzy-C Means—on raw data and apply linear (Principal Component Analysis) or non-linear (Variational Autoencoders), or a combination of PCA and VAE dimensionality reduction techniques. Results indicate that, for high-dimensional cognitive domains, a combination of PCA and VAE yields superior clustering quality. Conversely, in less high-dimensional domains, the VAE outperforms the PCA approach. In summary, the application of dimensionality reduction techniques demonstrates promising outcomes in student cognitive profiling, especially for data characterized by high dimensionality and heterogeneity. These findings have practical implications for advancing personalized learning and assessment experiences and enhancing our understanding of the intricate relationships within students' cognitive domains.

WEDNESDAY 3 JULY
Session 5.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

554. **Yoav Bergner (New York University), Yu Wang (New York University), Madhu Gopalakrishnan (New York University), Pranali Mansukhani (New York University), Ella Anghel (Boston College) On the Use of Large Language Models to Generate Novel Collaborative Problem Solving Items**

WEDNESDAY 3 JULY
Session 5.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research

803. Using Natural Language Models to Assess Cognitive Complexity of Academic Standards

Kevin O'Rourke
University of Massachusetts - Amherst

This study tests the effectiveness of a natural language processing model in rating the cognitive complexity of academic learning standards and to compare its performance with that of human subject matter experts. The primary objective is to evaluate ChatGPT's ability to apply Webb's Depth of Knowledge (DOK) framework and provide ratings comparable to those generated by experts. This study holds implications for the integration of AI technologies in educational assessment and offers insights into the reliability and validity of ChatGPT in the context of academic standard evaluation. The increasing role of AI in educational assessment prompts an examination of its capabilities in tasks requiring subjective judgment. While previous research has explored AI applications in education, this study distinguishes itself by specifically examining the performance of ChatGPT in rating learning standards, drawing attention to its potential contributions and limitations in comparison to human experts, as well as the challenges and implications of automating subjective tasks. The research questions for the study are: 1. To what extent can ChatGPT replicate the DOK ratings provided by human subject matter experts for academic learning standards? 2. How do ChatGPT's ratings/rationale compare with those provided by human SMEs? 5 subject matter experts (SMEs) were asked to give independent ratings of cognitive complexity (from 1 to 4, with one being the least complex, according to Webb's framework for Depth of Knowledge) for a set of academic learning standards. ChatGPT was asked to do the same for comparison. Further analyses include inter-rater reliability, comparison analysis with ChatGPT, and comparative qualitative analysis of rationale for ratings by SMEs and ChatGPT. Preliminary data/results show ChatGPT to provide DOK ratings consistent/comparable with those of human SMEs, but with possible problematic qualitative analysis results.

WEDNESDAY 3 JULY
Session 5.7
Topic: Construct or concept equivalence

84. Evaluation of item time effect on a computer- based selection test

Van Nguyen
Australian Council for Educational Research

uniTEST has been developed by the Australian Council for Educational Research to assist universities with the process of student selection. The test has been developed to assess the generic reasoning and thinking skills required to successfully complete studies at the tertiary level. In 2019, 5031 Danish candidates sat the test on computer-base: 4379 at campuses and 652 by remote Proctoring. The test consists of five components of 28 multiple-choice items each– Quantitative Reasoning, Critical Reasoning, Verbal-Plausible Reasoning, Scientific Reasoning, and Interpersonal Reasoning, in a mixed order. The aim of this study was to explore the time effect on the candidate performance and the test items. The amount of time that every test taker spent on each test item was recorded and then was taken count into the test and item analysis. Initial results showed that in general candidates spent more time on items related to science or quantitative reasoning, but less time on items related to humanity or verbal reasoning. There was a small correlation (statistically significant at 0.05) between item spending time and item position as well as item difficulty. Candidates tended to spend less time on items towards to the test end and more time on harder items. The correlation between item spending time and item fit statistics to the Rasch model (being used for scaling candidate scores) was not statistically significant in each of the test domains. Additionally, the pattern of item spending time is consistently similarly between the remote Proctoring and at campus groups, between males and females, and among the candidates' courses. Finding from this study would help for a consideration of keeping mix item structure for the test or separating items by individual domains as well as for giving useful information in construction of the tests by computer base.

WEDNESDAY 3 JULY
Session 5.7
Topic: Construct or concept equivalence

96. Cross-cultural variability in lay perceptions of mental illness in Germany: Measurement invariance over cultural groups, gender, age, and education

Marie Kollek
University of Hildesheim, Germany
Jesse Tse
University of Melbourne
Ronja Aileen Runge
University of Hildesheim, Germany

Lay perceptions of mental illness have a significant impact on personal experiences of mental illness as well as public attitudes towards individuals with such conditions. Recent research suggests that expanding the definition of mental illness to include a broader range of phenomena and lowering the threshold for considering unusual experiences as indicative of a disorder can reduce stigma and encourage help-seeking behavior. To quantitatively measure these lay perceptions, a new vignette-based scale, called the Concept Breadth-Vertical and Horizontal Scale (CB-V and CB-H), has been developed. Although studies have shown differences in CB-V and CB-H across cultural groups, it remains untested whether the CB-V and CB-H scale performs equally well across different cultures and socio-demographic characteristics such as gender, age, and education. This study aims to examine the measurement invariance in three distinct cultural groups within Germany: individuals of German, Turkish, and Chinese origin. Specifically, we are interested in exploring whether vignettes portraying somatic versus non-somatic symptoms, as well as intra- versus interpersonal symptoms, are similarly endorsed across cultures. Data collection is currently in the final stages, with samples comprising n=100 individuals of German, n=144 individuals of Chinese, and n=192 individuals of Turkish origin. Moderated Nonlinear Factor Analysis will be employed to test for invariance across cultural background, gender, age, and education. This method allows for the examination of invariance across interval and nominal scaled characteristics and enables the testing of both uniform and non-uniform differential item functioning. The outcomes of this study will not only enhance our understanding of the psychometrics of the CB-V and CB-H scale but also provide deeper insights into cultural variations in the conceptualization of mental health.

WEDNESDAY 3 JULY
Session 5.7
Topic: Construct or concept equivalence

620. Is It Worth to Use Constructed-Response Items in a Large-Scale Assessment Measuring Higher-Order Thinking?

Ozge Ersan, Ismail Cuhadar
Republic of Türkiye, Ministry of National Education

The use of constructed-response (CR) items along with multiple-choice (MC) items are in rise in Türkiye purporting to support students to think deeper, express their own ideas and extend their communication skills (Güneş, 2016; Kim & Cho, 2015; Ministry of National Education, [MoNE], 2023). In line with this purpose, a large-scale assessment called ABIDE (Testing and Evaluation of Academic Skills) is administered by MoNE to 4th, 8th and 10th grade students since 2016 to monitor their higher-order academic skills in Reading, Mathematics, and Science through real-life scenarios. The purpose of this research is to collect psychometric evidence for the use of CR items in such a large-scale assessment. Items coming from 4th grade Reading test administered in 2022 were examined to answer the research question: How do MC and CR items compare psychometrically? For this research question, (1) average item information functions of each item type for measurement precision, (2) average omit rates of items also considering item positions and item cognitive levels, (3) construct equivalence of MC and CR items in each form, (4) consistency of reading proficiency scores across two item types were examined. The analyses were conducted based on Item Response Theory using 12637 students' responses. The results show that, (1) CR items tended to provide higher measurement precision and (2) higher proportion of omit responses when compared to MC items. (3) Construct equivalence was observed for MC and CR items based on confirmatory factor analyses of each form. (4) Reading proficiency scores were highly correlated. Some psychometric evidence was collected and the results are promising for the use of CR items in a large-scale assessment to support higher-order thinking and communication skills in a centralized education system and testing culture familiar with MC items. Detailed results, discussion and references will be provided in final paper due to character restriction.

WEDNESDAY 3 JULY
Session 5.7
Topic: Construct or concept equivalence

694. How Cultural Cues Affect Bicultural Individuals' Personality Assessment Response Patterns: A Frame Switching Perspective

Patrick Lee, Charles Scherbaum
Baruch College & The Graduate Center, CUNY

A significant body of personality research has focused on cross-national comparison, but less attention has been devoted to within-country subcultural complexity – especially with regards to the growing population of bicultural individuals who subscribe to multiple cultural-value systems. Biculturals have been found to view situations under distinct frames of reference corresponding to their heritage and host cultures, and their attitudes and behavior can vary depending on which frame is activated at a given moment. The present study investigates whether such frame switching effects can be caused by elements within a formal personality assessment setting, and in turn affect biculturals' scores on the assessment. The moderating role of bicultural identity integration (BII) is also explored. East Asian immigrants in the United States were presented with various assessment-related cues (preceding tasks, administrator messages, organizational info, and test instructions) which were either more Asian or more American in their cultural content, and then completed a computerized inventory of the Big Five personality factors. Results showed that cultural framing and BII interacted to influence participants' scores, particularly for Agreeableness and Conscientiousness. Individuals with higher BII (i.e., more internal harmony between their two cultural identities) expressed personalities more aligned with the norms of their cued cultural frame, whereas lower-BII individuals exhibited contrast effects in the opposite direction. The findings highlight biculturals' multifaceted cultural selves as a potential contributor to measurement error in personality assessment scenarios. Organizations which use personality scores for decision-making purposes are recommended to review their test administration processes for unintended cultural content, and to be generally more cognizant of biculturals' ability to activate different cultural points of view from context to context.

WEDNESDAY 3 JULY
Session 5.7
Topic: Construct or concept equivalence

755. Using Q-matrices from CDM to inform the development of parallel test forms for the administration of diagnostic assessments in multiple languages

Sanet Steyn
University of Cape Town

The strength of a diagnostic assessment in educational settings lies not only in its accuracy but also in its ability to provide granular detail about a student's performance in a specific domain. The principles of Cognitive diagnostic modelling (CDM), and in particular, the use of Q-matrices to describe the attributes linked to individual test items, have been the subject of many investigations into the validation of diagnostic instruments and despite its strong reliance on human opinion, the framework they provide for the capturing test item specifications remain a valuable tool. Considering the high value that is placed on the comparability of test forms when using parallel instruments in high-stakes assessment contexts, the question of how one should go about ensuring parity between test forms that use different languages as the medium for assessment necessitates careful consideration of different frameworks that may facilitate assessment development. In the South African context, the National Benchmark Tests (NBTs), a set of assessments – in the domains of academic literacy (AL) and quantitative literacy (QL), as well as mathematics – used to measure student readiness for the academic demands of higher education study are a well-known high-stakes assessment administered in two languages: English and Afrikaans. This paper will explore the use of Q-matrices to inform the development of parallel test forms of the NBT AL and QL with a particular focus on the challenges in the classification process and how this is dealt with in this context.

WEDNESDAY 3 JULY
Session 5.8 SYMPOSIUM
Topic: Quantitative, qualitative, and mixed validation methods /Translation of tests, psychological assessment instruments and survey questionnaire/ Translation of tests, psychological assessment instruments and survey questionnaire

393. Seeing Beyond Scores: Eye Tracking as a Gateway to Understanding Cognitive Processes and Attentional Dynamics of Test Takers

Marlit Annalena Lindner
IWM, University of Tübingen
Bryan Maddox
Digital Education Futures Initiative, Hughes Hall, University of Cambridge & Assessment MicroAnalytics Ltd.

Eye-tracking provides insights into cognitive processes and attentional behaviors that are otherwise not observable. In educational assessment, these fine-grained process data can offer a deeper understanding of how test-takers comprehend materials, interact with them and how eye tracking data may even add incremental value to item responses and test scores. However, the potential of eye movement analyses is not exploited and deserves more attention. This pioneering symposium aims to contribute to the visibility of eye tracking research in the assessment community, by collecting four empirical studies that show promising applications and nuanced insights into test-taking behavior. The FIRST study investigated how multimedia-enriched, technology-enhanced test items (text and image vs. animations) affected test performance and attention in secondary students. Whereas student performance was not significantly affected (n=251), unique attentional behaviors could be identified via eye-tracking. (n=33). The SECOND study investigated n=10 secondary students' reading comprehension on digital screen and paper. Eye tracking data revealed that digital reading resulted in shallower processing, which has implications for the digitization of large-scale assessments. The THIRD classroom study with n=31 students shows how integrating eye-tracking with log data and machine learning enhances understanding of test-taker responses, offering a new view to support large-scale psychometric data. The FOURTH study with n=99 school children who solved multiple-choice items, shows that proficiencies can be estimated based on eye tracking data and deviations from typical viewing pattern may reveal disengaged behavior. Overall, the symposium underscores the potential of eye-tracking to enhance our understanding of test-taker interactions with items and their level of comprehension. It is also a decisive step to make this research more prominently seen in the educational assessment community.

WEDNESDAY 3 JULY
Session 5.8 SYMPOSIUM
Topic: Quantitative, qualitative, and mixed validation methods /Translation of tests, psychological assessment instruments and survey questionnaire/ Translation of tests, psychological assessment instruments and survey questionnaire

400. Investigating Response Models in Large-Scale Assessments: Refiguring Scale, Granularity and Diversity with Eye Tracking Studies

Bryan Maddox
Assessment MicroAnalytics Ltd

Investigations of assessment response processes using eye tracking techniques provide an important source of evidence for response model interpretations and their validation (Oranje, Gorin, Jia and Kerr, 2017). Their small grain-size provides a useful counterpoint to large-scale psychometric data, to investigate the way that test takers approach and find solutions to assessment tasks (Kane and Mislevy, 2017). However, the scale of eye tracking studies mean that they are thought to lack statistical power and discrimination. As a result, such 'direct probes' (Messick, 1995), while 'illuminating' (Messick, 1995, p743), have often been presented as peripheral, secondary sources of 'para-data'. Furthermore, with the rise of log data (which offers both granularity and scale), we might ask what additional value eye tracking studies offer? To answer this question the paper refigures the contribution of eye tracing studies in large-scale assessments. This paper develops a new 'interpretation and use argument', and associated 'warrants' (Kane, 2016), to highlight the valuable contribution of eye tracking studies in large-scale assessments. The paper provides a case study involving a classroom-based eye tracking study (n=31) of secondary school mathematics assessments in France. The study was conducted with DEPP, as the Ministry of Education in France. It demonstrates how eye tracking studies can support large-scale inferences, when they are integrated with log data analysis and the use of machine learning techniques. Furthermore, it shows how the granularity - precision and accuracy of screen mounted eye trackers and integrated video data can be used to differentiate and understand the sources of diversity in test taker responses, and to compensate for the inherent partiality of log data. In conclusion, the paper argues that eye tracking studies integrated with log and psychometric data can generate powerful insights into the diversity of test responses.

WEDNESDAY 3 JULY
Session 5.8 SYMPOSIUM
Topic: Quantitative, qualitative, and mixed validation methods /Translation of tests, psychological assessment instruments and survey questionnaire/ Translation of tests, psychological assessment instruments and survey questionnaire

491. Potential Contributions of Eye Movement Parameters in Measuring Ability via Multiple-Choice Items

Marlit Annalena Lindner
IWM, University of Tübingen
Esther Ulitzsch
CEMO, University of Oslo
Gabriel Nagy
IPN, Leibniz-Institute for Science Education, Kiel

Eye movement measures can provide valuable insights into examinees' solution processes and their preferences for different answer options in educational multiple-choice items, known as the gaze bias effect. This suggests the potential for inferring examinees' task-specific knowledge based on viewing patterns during their item solution process. However, it is unclear how eye movement patterns could be aggregated to complement test score information in educational assessments. In this study, we aimed to construct an index based on examinees' relative attention distribution to the answer options in multiple-choice items to investigate the possibility of assessing student proficiencies and (dis-)engagement solely based on eye movements. We also explored whether the data could provide incremental information beyond item responses (correct vs. wrong). Our dataset comprises 99 school students who solved 18 single-choice science items with four answer options each. We analyzed eye tracking fixation data focused on the four individual answer options (i.e., areas of interest) at the item level. The answer option correctness and relations between option fixation times were considered in the index construction to predict individual test scores at the item and student level. The item-level index differentiated well between correct and incorrect responses. The aggregated index correlated strongly with students' test scores. The results indicate that test scores and eye tracking indices measured the same construct with a similar reliability. Furthermore, the pattern of students' attention distribution allowed to identify examinees that were conspicuous of guessing and showing disengaged behavior. Implications for educational assessment practice, limitations, and challenges of our approach will be discussed.

WEDNESDAY 3 JULY
Session 5.8 SYMPOSIUM
Topic: Quantitative, qualitative, and mixed validation methods /Translation of tests, psychological assessment instruments and survey questionnaire/ Translation of tests, psychological assessment instruments and survey questionnaire

421. Paula Lehane (Dublin City University) Technology-Based Assessments (TBAs): Using eye movement data to understand test-taker attentional behaviour

WEDNESDAY 3 JULY
Session 5.8 SYMPOSIUM
Topic: Quantitative, qualitative, and mixed validation methods /Translation of tests, psychological assessment instruments and survey questionnaire/ Translation of tests, psychological assessment instruments and survey questionnaire

512. The smell of paper or the shine of a screen? Students' reading comprehension, text processing, and attitudes when reading on paper and screen

Ragnhild Engdal Jensen, Astrid Roe, Marte Blikstad-Balas
University of Oslo

With the increasing prevalence of digital devices such as smartphones, tablets and e-readers, more and more reading is happening in digital formats – also in the educational context. The present study focuses on lower secondary school students and their reading comprehension and attitudes toward reading on paper and on screen. The study uses an innovative methodological approach. Eye-tracking technology was used to observe the reading of carefully sampled students (n=10) at different reading comprehension levels. The students read a selection of texts and answered questions from the Norwegian national reading assessment on comparable versions on paper and on screen. By analyzing eye movement data (reading transitions), including more than 25000 fixations, in combination with text comprehension outcomes, students´ cued retrospective reporting, and interview data we have obtained detailed and comprehensive data on students' reading comprehension, reading behavior, and reading experiences across different mediums. A key result is that reading on screen leads to more shallow processing and can hinder reading comprehension. Importantly, our results from the students' cued retrospective reporting of their eye tracking, showed that they were unaware of their reading behavior and didn't reflect much on reading in different mediums. These findings have implications for the increasing shift to digital learning environments, including the digitization of large assessments, in the educational context. It is important to recognize the difference between reading processes, and policymakers and practitioners cannot assume that these processes are the same across individuals and different delivery modes.

WEDNESDAY 3 JULY
Session 5.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

74. Can an adjustment index address the norming challenges in testing in South Africa? Investigating the use of an index to adjust scores

Yaseen Ally, Justin August
Nelson Mandela University

With the changing landscape in South Africa to a full democracy, increased research has been undertaken in the psychometric field on local and national normative studies regarding various assessment measures. While norms for populations have been developed in South Africa, these studies are based on race and gender predominantly, without considering the effects of socio-ecological factors on test performance. In this presentation, we draw on Bronfenbrenners socioecological framework and argue that test-taker performance is a culmination of ability and exposure. In doing so, we position the need for more criticality when interpreting test scores in contexts where diverse ethnicities, languages and culture exists. In addition, we introduce the role an adjustment index might have within the field of testing and assessment.

WEDNESDAY 3 JULY
Session 5.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

219. Fit for purpose or off the mark: Pirls 2016 in South Africa

Heather Leigh Kayton
University of Oxford
Yasmine El Masri
Ofqual

This research is situated within the context of widening participation in international large-scale assessments (ILSAs) and the resulting tensions between accommodating an increasingly diverse group of countries and ensuring valid, reliable, and comparable results. Despite consistently low performance compared to other countries, developing countries such as South Africa often rely on ILSAs to reveal insights into their education system through the lens of student performance. However, there has been considerable evidence that ILSAs fall short when it comes to assessing students from lower-performing countries. This research evaluates whether PIRLS Literacy is able to meet the needs of the South African context, which are complex in terms of linguistic diversity, wide attainment gaps and extreme inequality. A subset of 4,150 Grade 4 students tested in Afrikaans (n=1,228), English (n=2,089) and Sepedi (n=898) was selected from the PIRLS 2016 national dataset for South Africa. These groups were chosen based on their relative overall performance to enable comparison across higher and lower performing language groups. The study applied Item Response Theory (IRT) models (2PL and GPCM) to examine the relationship between item difficulty and student ability. Model fit comparisons, checks of dimensionality and local dependence, and graphical examinations of item and student parameters were conducted. The findings show that more than one third of the items did not fit the students in this dataset. Furthermore, more than half the items were too difficult to provide any information for the below average students in all three groups, and only 3% of the items were appropriately targeted for below average students in the Sepedi group. Overall, the evidence suggests that PIRLS Literacy may not be a valid measure of student reading achievement in South Africa, particularly when comparing performance across language groups, and the information drawn from it may not be meaningful.

WEDNESDAY 3 JULY
Session 5.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

391. Impact of moving high-stakes assessments on-screen on students in England

Yasmine El Masri, Jo Handford
Ofqual

With greater adoption of digital assessments globally, England is examining the opportunities and risks of moving high-stakes assessments from pen and paper to the computer screen. The literature outlines many benefits of digital assessments, including better student experience, higher levels of engagement and performance as well as reduced levels of cognitive load compared to paper-based assessments. Moreover, the on-screen mode could make assessments more accessible, especially for students with special educational needs (SEND). While the benefits outlined are certainly attractive, many of the claims are not supported by robust evidence and much of the research has been carried out in the context of higher education and low-stakes assessments. In addition, the evidence available is at best mixed as the levels of performance, cognitive load and engagement appear to be influenced by various factors including assessment design, subject matter and students' characteristics, abilities, familiarity with and access to digital devices. Given the significant potential impact of greater adoption of on-screen high-stakes assessment on students, the Office of Qualifications and Examinations Regulation (Ofqual) and the Department for Education have undertaken research to better understand stakeholders' perspectives on the nature and extent of the impact of on-screen assessments on different groups of students (age, socioeconomic background, special educational needs) in the context of English high-stakes examinations. Interviews with students, parents, teachers and subject matter experts (principals, SEND coordinators, psychologists) selected from diverse backgrounds were carried out. Key themes include concerns over the rate of implementation, system readiness as well as fairness given the existing inequities in the system. With Ofqual prioritising the impact of its regulations on students, this research may inform regulatory approaches Ofqual adopts in the future.

WEDNESDAY 3 JULY
Session 5.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

436. Decolonising Outcome Measurement with Kaupapa Māori Psychometrics: A systematic review and methodological quality appraisal of health and wellbeing measures for Māori

Richard Siegert
Auckland University of Technology
Margaret Sandham
Massey University
Maree Roche, Melissa Carey
University of Auckland
Rebecca Jarden
University of Melbourne

Background: Dimensions of health and wellbeing relevant to indigenous groups must be identified and represented in outcome measures to address the consistent pattern of inequitable health outcomes experienced by indigenous populations. Ensuring outcome measures used with indigenous populations have content validity is the first step when assessing progress towards equity in health outcomes. Objectives: To review the approaches Māori researchers have used to develop health and wellbeing outcome measures, and describe how measure development has aligned with COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) criteria. Design: Systematic review and methodological quality appraisal. Data sources: CINAHL (Ebsco), EMCARE, MEDLINE, psycINFO, SCOPUS electronic databases were searched in May 2022 with no date limiters. Publications were considered for inclusion if published in English or Te Reo and reported Māori health and wellbeing outcome measure development. Review methods: Key search terms identified studies developing and validating Māori health and wellbeing outcome measures. Studies were appraised using COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN). Results: Seven publications reported the development of outcome measures and a further seven related to validation of the measures. All health and wellbeing measure development studies or content validity studies were categorised as "Inadequate" using the COSMIN criteria so measurement properties were not appraised. Conclusion: Psychometrics often neglect contextual and cultural influences and in contrast, Kaupapa Māori embraces these aspects through a qualitative approach. Combining the strengths of both approaches yields culturally relevant outcome measures, promoting equity in health and well-being assessment for Māori people. Keywords: health; Māori; measure; wellbeing; instrument; outcome measurement; COSMIN

WEDNESDAY 3 JULY
Session 5.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

756. Towards a Justice-oriented Approach to Assessment: Unpacking Religious Bias in English Language Assessments from Muslim Teachers and Students

Sheila Lallmamode
AILLA Lab

This research delves into the often-neglected realm of religious biases in language assessments, paralleling the oversight of biases related to culture, identities, and race. While guidelines exist to circumvent 'objectionable' content in language tests - such as racism, migration, war, ableism and murder - addressing the broader historical and socio-political context and its impact on underrepresented communities remains an understudied area. Situated within a justice-oriented approach to assessment, this study scrutinizes religious bias in popular English language tests - namely IELTS and TOEFL - as perceived by Muslim lecturers and university students in Saudi Arabia. The primary objectives encompass: (1) identification of religious bias elements;(2) examination of how Muslims perceive sensitive cultural exclusion and inappropriateness in the tests; and, (3) assessment of how justice-oriented reading texts tailored to Muslim test-takers are perceived. Adopting a mixed-method approach, teachers and students responded to a questionnaire addressing religious bias elements, while focus group discussions were conducted with ten English language lecturers and ten university students. Sample materials were sourced from official test preparation books and websites of two major English language tests, the IELTS and TOEFL. Insights derived from the study indicated that materials in the standardized tests inadequately addressed the historical, racial, cultural, religious, and social complexities shaping Muslim test-takers. The implications underscore the imperative need to personalize and regionalize assessments that cater to the unique experiences of Muslim students in the Middle East and beyond. This study advocates moving beyond 'sanitized' tests that strive to avoid offense and emphasizes the profound impact of assessments on underrepresented communities. This study is a call to action for language test developers to genuinely reflect the diverse voices of the world.

Round table by Beatrice Rammstedt, GESIS University of Mannheim (Germany), Ana Villar, Meta (UK) and Bruno D. Zumbo, University of British Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

565. Exploratory Graph Analysis VS Exploratory Factor Analysis: A comparison of two methods for identifying the dimensional structure of the SHSS.

Alberto Sanchez-Fernandez-Quejo, Carla Perez-Guerra, Marco Innamorati, Alejandro de la Torre-Luque
Universidad Complutense de Madrid, Spain

Conceptual Framework: The Suicidal History Self-Rating Screening Scale (SHSS) assesses suicidal ideation over the past year and throughout the life before the last year. Satisfactory properties have been found for the original, the Italian version, using a sample from clinical (hospital) settings. Further research is needed to provide data on psychometric performance in the general population. Moreover, psychometrically valid tools should be developed or translated into the Spanish population, due to the paucity of suicide risk measures for Spanish population. Objectives: This study aims to provide some psychometric data of the Spanish SSHS by studing its internal structure. Two different methods were used: Exploratory Factor Analysis (EFA) and Exploratory Graph Analysis (EGA). The sample consisted of 400 Spanish young adults (M=21.35 years old; SD=3.68). Methodology: The primary approach involved utilizing EFA and EGA to detect the dimensional structure. Results: EFA revealed a two-factor structure, distinguishing between suicidal ideation and behavior, with some cross-loading evidence for the item 9. In contrast, EGA revealed a single dimension for each vital moment and five dimensions when considering both, distinguishing between the vital moments in terms of suicidal behaviour despite grouping some ideation items. Implications: Two differentiated factors were revealed, standing for the two main forms of suicide-related behavior (i.e., ideation and behaviors). EFA and EGA may be considered complementary methods for estimating dimensional structures, but they exhibit differences in both the process and results. These distinctions will be thoroughly discussed in this work, incorporating insights from SHSS.

Round table by Beatrice Rammstedt, GESIS University of Mannheim (Germany), Ana Villar, Meta (UK) and Bruno D. Zumbo, University of British Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

216. Adaptation and Validation of the "Detail and Flexibility Questionnaire (DFlex)" in a Spanish Population

Alicia Georghiades
ITA Salud Mental, Barcelona, Spain
Esteve Montasell
ITA Salud Mental y Autonomous University of Barcelona, Barcelona, Spain
Ester Serrano
Nerea Lopez Team, Child Psychology Centre, Castelló, Spain
Beatriz Lanceta, Jordi Alabernia, Antoni Grau
ITA Salud Mental, Barcelona, Spain

Theoretical framework: Eating Disorders (EDs) can cause complications and may be related to a disturbance of central coherence and rigid behaviours arising from executive dysfunction. This can be associated with an inflexible and overly detail-specific cognitive style or, on the contrary, to impulsive behaviours. The Detail and Flexibility (Dflex) questionnaire allows for these concepts to be measured but, as of yet, has not been made available in Spanish. Objectives: The purpose of this study is to explore the psychometric properties of the questionnaire in a Spanish population. In addition, we want to assess whether the questionnaire provides information on the cognitive profile of the adult population with different EDs and to see whether it discriminates between the community sample and clinical population. Sample: A minimum of 200 Spanish adult inpatient and day hospital patients with an eating disorder diagnosis are expected to participate. They will be recruited from the ITA Salud Mental Health Eating Disorders Unit. Methodology: The translation and validation process followed the guidelines as outlined in the 'International Test Commission'. Three independent translators were involved in said process. Results: We are currently in the pilot study, in which the questionnaire has been administered to 35 patients. Once data collection has ceased, we will administer it to additional patients to ensure its validity. Implications: To our knowledge, there are no self-report measures in Spanish that assess the aforementioned domains. Self-report measures can complement existing objective measures, thus uncovering a lack of awareness on the part of people affected by eating disorders and alterations in the area of cognitive rigidity. Low cognitive flexibility is an endophenotype present in eating disorders and may have important implications for the treatment of these pathologies. Future research may focus on community samples in order to explore cut-off points.

Round table by Beatrice Rammstedt, GESIS University of Mannheim (Germany), Ana Villar, Meta (UK) and Bruno D. Zumbo, University of British Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

134. Spanish adaptation of the Hierarchical Taxonomy of Psychopathology-Self Report (HiTOP-SR): cultural issues across different Spanish-speaking countries.

Ana María de la Rosa-Cáceres, Carmen Díaz-Batanero
University of Huelva (Spain)
Deisy Gonzalez-Zapata, Melissa Briones
University of North Texas (USA)
Jennifer Callahan
University of Texas at Dallas (USA)
Nazaret Fresno Cañada
University of Texas at Rio Grande Valley (USA)
Roman Kotov
Stony Brook University (USA)
Leonard Simms
University at Buffalo (USA)
B. Villalobos
University of Texas at Rio Grande Valley (USA)
Camilo J. Ruggero
University of Texas at Dallas (USA

Over the past few decades, several dimensional models have been developed to overcome the limitations of categorical models in psychopathology. One of the most promising dimensional models is the Hierarchical Taxonomy of Psychopathology (HiTOP), which has shown important advances in clinical and research contexts in recent years. Based on the evidence found in the literature, the HiTOP model reorganizes the classification of symptoms into a hierarchy of six levels ordered from most to least broad: superspectra, spectra, subfactors, syndromes, homogeneous symptom components/maladaptive traits, and symptoms. To date, this model's research progress has been sustained on different aligned instruments, which are already available but do not completely cover the whole model. The HiTOP-Self Report (HiTOP-SR) is the first instrument specifically designed to assess the symptomatic components of all spectra. Premier results on

English-speaking samples are being launched. The aim of this study is to present the adaptation process followed to reach an agreed version for a Spanish version of the HiTOP-SR, ready to be used in different Spanish-speaking countries. The adaptation considers linguistic and cultural differences, following the ITC guidelines. It includes grammatical, syntactic, lexical, and semantic considerations. Preliminary results of the adaptation process and examples of some culturally sensitive items are presented.

Round table by Beatrice Rammstedt, GESIS University of Mannheim (Germany), Ana Villar, Meta (UK) and Bruno D. Zumbo, University of British Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

652. Why is translating and analyzing the internal structure of scales not enough to certify their suitability in other cultures? An example with the Psychological Functioning Scale.

Ana Paula Noronha, Ligia Santis, Monique Guimarães, Ana Paula Cavallaro, Leila Couto
Universidade São Francisco

Research has shown that simply translating scales from their original language into another may not be enough. In this study, we present the detailed process of translating the Positive Psychological Functioning Scale (PPF) from Spanish to Brazilian Portuguese. Positive psychological functioning refers to personality traits that can be learned, enabling individuals to adapt, develop personally, and achieve what is important to them. The goal of this study is to describe the translation and adaptation process of the PPF for use in Brazil. The translation process involved four judges that were instructed to read the definitions of the 11 psychological resources and their respective items, and translate them while considering cultural, and scientific aspects. The results of the translations were then discussed by an expert committee, who selected the translated options that accurately reflected the intended meaning of each item within its corresponding dimension. After, the items were evaluated for comprehension by the participants. Eleven participants, aged between 18 and 54, participated in individual structured interviews. Next, the scale was analyzed by nine judges who assessed the agreement between them regarding the distribution of the items into the 11 categories of the original scale, as well as the language clarity, practical relevance, and theoretical relevance. Additionally, the scale was back-translated by four bilingual translators. The committee discussed decisions about potential changes to wording and the age group/target audience of the instrument. The final version of the scale was submitted to the original authors for their review and approval. This study highlights the importance of not relying on translation and analyzing the internal structure of scales when assessing their suitability in different cultures. Cultural adaptation, as well inputs from the participants and experts, are necessary to ensure the validity of the scale within the contexts.

Round table by Beatrice Rammstedt, GESIS University of Mannheim (Germany), Ana Villar, Meta (UK) and Bruno D. Zumbo, University of British Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

708. Focus on selection versus development for increased job performance. Four decades of research and a Monte Carlo simulation of individual differences and job attitudes.

Andreea Corbeanu, Dragos Iliescu, Andrei Ion
University of Bucharest

Through a systematic meta-analytic review and Monte Carlo simulation, this study explores the differential validity of individual differences (usually approached through organizational selection practices) and job attitudes (usually approached through organizational interventions), when it comes to predicting job performance. Based on a meta-analytic correlation matrix between all study variables, we employed a Monte Carlo simulation to generate a dataset reflecting 10k fictitious companies, each with its own particularities. Based on this dataset, we deployed a series of regressions, relative predictor weight analyses, and utility analyses. Our results suggest that a total of 38% of the overall variance in task performance and 48% in contextual performance, respectively can be accounted for by a model including both individual differences and job attitudes (R2 = .38 and .48, respectively). Cognitive abilities (General Mental Ability - GMA) had the highest predictive power in both models, = .63 (C.I. 90 = [.48, .77]) for task performance, and = .45 (C.I. 90 = [.32, .59]) for contextual performance. The utility analyses, suggest that interventions in job attitudes are a net loss when their effects are equal or below d = .20 and a cost of $3,000 or more per person; break-even is achieved only for interventions where the cost per employee is below $900. Conversely, investments in selection will represent a net loss only when the gains in performance are below d = .60, at a cost of $4,700 per hire. Theoretical and practical implications of these findings are explored within the discussion section.

(Germany), Ana Villar, Meta (UK) and Bruno D. Zumbo, University of British Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

383. Psychometric Properties of the Spanish-SAM Motives Measure (S-SMM) among Young Adults who use cannabis.

Lucía Vélez-Pérez, Bella María González Ponce
Department of Clinical and Experimental Psychology, University of Huelva (Spain)
Angelina Pilatti
Institute of Psychological Research, IIPsi-CONICET-UNC, National University of Córdoba (Argentina)
José Carmona-Márquez
Department of Clinical and Experimental Psychology, University of Huelva (Spain)

Background: Simultaneous alcohol and cannabis use (SAM), so that their effects overlap, is associated with greater negative consequences compared to non-SAM use. Motives for substance use are identified as strong predictor of alcohol and cannabis use and its consequences, and measures with robust psychometric properties are available for both substances. Recently, the SAM motives scale (Patrick et al., 2018) and its short version (Conway et al., 2020) have been developed; but no Spanish version of this or any other SAM motives measure is available. Objective: We aim to provide a Spanish adaptation of the SAM motives scale by Patrick et al. (2018), and examine its psychometric properties in community young adults reporting cannabis use. Method: Within a longitudinal project (Psicocann), targeted sampling was applied to access a sample of 612 past-month cannabis users, of whom 479 reported engaging in SAM use (Mage=21.01, SD=2.14; 36% female). After the adaptation-translation process of the items by a group expert, participants completed the Spanish-SAM Motives Measure (S-SMM) and measures for SAM use, cannabis use motives and negative consequences of alcohol and cannabis. Confirmatory factor analysis was used to assess the fit of the S-SMM to the original correlated four-factor structure. Results: Our findings showed acceptable fit (CFI=.95, TLI=.95, RMSEA=.07, SRMR=.08), and S-SMM scores were positively correlated, as theoretically expected, with SAM use, motives for cannabis use, and negative consequences of alcohol and cannabis. Cronbach´s Alpha were: conformity=.74; positive effects=.88; calm/coping=.84; social=.70. Conclusions: Our results support the utility of S-SMM among young adult SAM users, and could be useful for inform interventions aimed at minimizing SAM-related harms. This work was supported by the Agencia Estatal de Investigación (Ministerio de Ciencia e Innovación, Spain) under Grant Number PID2020-118229RB-I00 (PI: Fermín Fernández Calderón).

(Germany), Ana Villar, Meta (UK) and Bruno D. Zumbo, University of British Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

577. **The Development of the SEL-90 Test**

Butucescu Andreea, Iliescu Dragos
University of Bucharest

The SEL-90 Test comprises 90 questions assessing social and emotional competencies, along with 12 questions for social desirability. Individuals being evaluated respond to statements regarding their behavior, skills, preferences, attitudes, and self-perceptions, indicating how often these apply to them in general. The test typically takes about 35 minutes but may be longer for younger children or those with reading difficulties. It is designed for individuals aged 7 to 18. Scores are calculated for both broad and narrow competencies and are expressed in raw scores, T-scores, and percentiles. The development of the test involved writing, refining, and selecting questions, with subsequent adjustments. Items are assessed using a Likert scale ranging from 1 (Very rarely true about me) to 4 (Very often true about me). The test measures social and emotional competencies in children and adolescents and has been developed and validated through multiple stages.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

361. Development of the Index of Intensity of Violence Against Women (IIVM) in Peru.

Carlos Renzo Rivera Calcina, Rodolfo J. Castro Salinas, Walter L. Arias Gallegos
Universidad Católica San Pablo, Perú
Mitchell Clark
Mount Royal University, Canada
Carlos Renzo Rivera Calcina, Rodolfo J. Castro Salinas, Walter L. Arias Gallegos
Universidad Católica San Pablo, Perú
Mitchell Clark
Mount Royal University, Canada

This research describes the psychometric properties of the Index of Intensity of Violence against Women (IIVM) taken from the Demographic and Family Health Survey (INEI, 2023) that is applied annually by the National Institute of Statistics and Informatics to women between 15 and 49 years throughout the entire territory of Peru. The 12 items of the survey that assess domestic violence were considered, in a sample of 18,503 women. First, an exploratory factor analysis (EFA) was applied to assess its internal structure as well as the McDonald's Omega Test to estimate reliability. Then, using the criteria of the judges, the intensity index of the violence was calculated and the scales for qualification for the levels of violence were calculated. It was determined that the IIVM has a one-dimensional structure with an optimal reliability index (= .859), and three levels for its rating. It was estimated that based on the proposed scales, only 14.9% of the sample had been victims of domestic violence.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

206. An Analisis of the Partner's Perceived Responsive Scale in Mexico.

Carolina Armenta Hurtarte
Universidad Iberoamericana
Pablo Tonathiu Salcedo Callado
Centro de Integración Juvenil, A. C.
María Bárbara Rivero Puente
TECNOLOchicas

The Partner's Perceived Responsive Scale, rooted in the Interpersonal Process Model of Intimacy, posits that self-discovery and responsiveness within a couple contribute to intimacy and are linked to crucial aspects of relational dynamics. Given the growing relevance of this model, this study aims to investigate its applicability within a collectivist context, specifically in Mexico. Our sample comprised 160 participants, all aged 18 and above, involved in romantic relationships, with 70% identifying as women. Through confirmatory factor analysis, we obtained favourable adjustment indices, affirming the scale's validity. Our findings highlight the profound impact of shared activities on perceived responsiveness and overall satisfaction within romantic partnerships. This exploration enhances our understanding of the cross-cultural relevance of the Partner's Perceived Responsive Scale, shedding light on its implications for relationship dynamics in collectivist societies like Mexico.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

604. Design and development of a Gamified Test to assess Critical Thinking in Personnel Selection Contexts.

Virginia Arranz, Sonia Rodríguez, Beatriz Lucía, David Aguado
Instituto de Ingeniería del Conocimiento

Technological development has led to a significant evolution of assessment techniques in personnel selection contexts (e.g., gamification). However, as some authors point out, an aspect that has not benefited from this technological development is the measurement of new constructs of interest, such as competencies that are positioned as crucial for professional performance in VUCA environments. One of these competencies is critical thinking, considered as a higher form of reasoning, a cross-cutting skill in the educational context, and an essential cognitive resource in the professional realm. The aim of this work is to present the design of a gamified assessment test for evaluating Critical Thinking, as well as to provide initial evidence of its psychometric properties. To achieve this, the paper describes the design and development process of the test and the initial psychometric characteristics obtained in a first pilot study (N=90) conducted with psychology students and employees from a retail company. Results show that, based on the developed test, it is possible to infer the critical potential of individuals, not only from the perspective of their cognitive abilities but also from their disposition and effort to make reasoned decisions. The study contributes value regarding the evaluation of a possible predictor in the field of personnel selection, such as Critical Thinking. It also supports the need for psychometric development to analyze new assessment methodologies based on gamification. The use of gamification, breaking away from the traditional composition of classic tests, offers valuable possibilities in the field of personnel assessment. Among these, the possibility of simulating assessment scenarios close to reality and promoting the motivation and commitment of the individual to task execution, leading to higher quality information obtained about the evaluated subject.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

730. Psychometric properties of the Screening Test for Reading and Spelling Difficulties for Lithuanian speakers in second grade.

Dovile Butkiene
Vilnius University, Institute of Psychology, Lithuania
Reda Gedutiene
Klaipeda University, Department of Psychology, Lithuania
Lauryna Rakickiene, Kestutis Dragunevicius, Grazina Gintiliene
Vilnius University, Institute of Psychology, Lithuania

Identification of students at-risk for reading and spelling difficulties in Lithuanian primary schools was limited due to the lack of language based and norm-referenced screening tools. This prompted the development of the Screening Test for Reading and Spelling Difficulties, designed as a group-administered test in the second grade for speakers of Lithuanian. The test is based on the phonological deficit hypothesis of reading and spelling difficulties and consists of five subtests that measure letter recognition, word recognition and decoding, sentence reading, sentence copying, and phonological awareness. The scores of the three subtests (Word Chains, Sentence Chains and Sentence Copying) are used to calculate a composite Literacy Scale score. The aim of this study was to examine the psychometric properties of this test. A representative sample consisted of 252 Lithuanian second-graders (126 boys, 126 girls) aged 7:9–9:3 years. Sample size for test-retest reliability was 54 children. In addition, 46 second-graders with diagnosed reading and/or spelling disorders were tested. Split-half reliabilities for subtests ranged from 0.87 to 0.96, for composite Literacy Scale score was 0.97. Retest reliabilities (after 3–4 months) for subtests ranged from 0.60 to 0.88, for composite Literacy Scale score – 0.90. Inter-rater reliability for the Sentence Copying subtest was 0.98. Moderate to strong correlations (0.44–0.71) were found between subtests, composite Literacy Scale scores, and perceived pupils' reading and writing difficulties (teacher's ratings). Significantly worse performance of subtests in second-graders with diagnosed disorders (Cohen d ranged from 0.61 to 1.81) supports the concurrent validity of the test. To conclude, the present study gives evidence of adequate psychometric properties of the Screening Test for Reading and Spelling Difficulties for Lithuanian speakers and shows the usefulness of the test as a screening instrument for identification purposes.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

180. The Zarit Burden Interview (ZBI-12): A Reliability Generalization Meta-analysis.

Elena Cejalvo Herraiz
Department of Basic Psychology, Faculty of Psychology and Speech Therapy, Universitat de València, Spain
Júlia Gisbert-Pérez, Laura Badenes-Ribera
Department of Methodology of Behavioral Sciences, Faculty of Psychology and Speech Therapy, Universitat de València, Spain
Manuel Martí-Vilar
Department of Basic Psychology, Faculty of Psychology and Speech Therapy, Universitat de València, Spain

INTRODUCTION The Zarit Burden Interview (ZBI) is a tool used to assess caregiver burden, addressing physical, emotional, social, and economic aspects. OBJECTIVES Conduct a reliability generalisation meta-analysis study to estimate the average reliability of the ZBI-12 version (Bédart et al; 2001) and assess the degree of heterogeneity of the reliability coefficients in various samples and contexts. METHOD A reliability generalisation meta-analysis was performed following REGEMA guidelines (Sánchez-Meca et al., 2021). Web of Science (WoS), Scopus, Pubmed and PsycArticles were searched for studies. Of the 33 studies identified that applied the ZBI-12 version, only 20 provided reliability estimates using Cronbach's alpha coefficient and were included in the meta-analysis. For statistical analysis, Cronbach's alpha coefficients were transformed according to Bonett's (2002) formula to normalise their distributions. They were then returned to the Cronbach's alpha coefficient metric to facilitate the interpretation of the results. In the statistical analysis, a random effects model was applied to estimate the mean reliability coefficient and its 95% confidence interval using the method proposed by Hartung and Knapp (2001). RESULTS The mean reliability of the ZBI-12 total scores assessed as internal consistency using Cronbach's alpha coefficient was 0.865 (95% CI [.844,.884]). There was a large heterogeneity among the reliability coefficients. CONCLUSION According to psychometric theory, the ZBI-12 shows outstanding reliability, which makes it suitable for exploratory purposes in both academic and clinical practice. Keywords: caregivers, burden, reliability, meta-analysis.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

336. Psychometric properties and normative data of the Spanish S-UPPS-P Impulsive Behavior Scale in adolescents.

Esteve Montasell-Jordana
Departament de Psicología Clínica i de la Salut, Universitat Autònoma de Barcelona, Barcelona, Spain
Eva Penelo
Departament de Psicobiologia i de Metodologia de les Ciències de la Salut, Universitat Autònoma de Barcelona, Barcelona, Spain
Laura Blanco-Hinojo
MRI Research Unit, Department of Radiology, Hospital del Mar, Barcelona, Spain
Anna Soler, Beatriz Lanceta, Alba Ollé
Neuropsychology Department, ITA Salud Mental, Clariane Group, Barcelona, Spain
Jesús Pujol
MRI Research Unit, Department of Radiology, Hospital del Mar, Barcelona, Spain
Joan Deus
Departament de Psicología Clínica i de la Salut, Universitat Autònoma de Barcelona, Barcelona, Spain

Conceptual framework: Assessing impulsivity in adolescents is challenging due to limited age-appropriate tools. The UPPS-P impulsivity model has become a valuable tool. The short UPPS-P is a 20-item self-report instrument with five subscales: Negative Urgency, Lack of Premeditation, Lack of Perseverance, Sensation Seeking and Positive Urgency. Objective: To examine the psychometric properties of the short Spanish version of the UPPS-P scale and to provide normative data for adolescents. Sample: The study included 9733 adolescents aged 11-19 years with 8944 participants from 66 Spanish high schools and 789 patients from a mental health clinic. Methodology: Internal structure and measurement invariance across gender and age of S-UPPS-P were analyzed with Confirmatory Factor Analysis (CFA) and Exploratory Structural Equation Modeling (ESEM) with target rotation, both for categorical indicators with weighted least squares means and variance (WLSMV). Internal consistency reliability of the S-UPPS-P scale scores was assessed with the omega coefficient. Pearson's correlation coefficients were used to examine the convergent validity with the Barrat Impulsiveness scale (BIS-11-A) global score. Gender and age differences were evaluated with a two-factor mixed ANOVA. Results: CFA revealed that the 5-factor model was inadequate but it demonstrated satisfactory fit when employing ESEM approach (CFI and TLI ≥ .97, RMSEA = .035). Measurement invariance across gender and age was achieved. Internal consistency reliability showed modest to satisfactory values (omega =

.67 to .82). All S-UPPS-P subscale scores except Lack of Perseverance and Sensation Seeking were moderately correlated with the BIS-11-A total score (r = .47 to .59). No relevant gender or age differences were found. Norms were calculated using T-scores and percentile ranks for the whole sample. Implications: We provide standardized norms that can be used for scholar counseling, daily clinical and research practice.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

712. Critical consciousness in sport scale (CCSS): Development and psychometrics properties.

Evandro Peixoto
São Fracisco University
Martin Camiré
University of Ottawa Canada

Current theoretical proposals have suggested the potential of sport to help athletes develop their critical consciousness (CC) in relation to the forces that shape society as well as the systems of privileges and deprivations that influence access to sport. The present study proposes the initial development of the Critical Consciousness in Sport Scale (CCSS), offering initial validity evidence based on test content, internal structure, relations with other variables, and reliability. Based on Paulo Freire, CC has three components: Critical Reflection; Political Efficacy, and Critical Action. An initial pool of 50 items was developed around four central themes in sport: (a) racism (e.g., Black people are victims of racism in sport); (b) gender inequality (e.g., I can contribute to fostering gender equality in sport); (c) sexism (e.g., I support on social media the LGBTQ community in sport); and (d) socioeconomic inequality (e.g., I participate in activities that promote social equality in sport). The items were submitted to the evaluation of three expert researchers/PhDs in the psychology, pedagogy, and sociology of sport who confirmed the suitability of the items' content in terms of practical relevance and theoretical adequacy. The sample was comprised of 263 Brazilian psychology and physical education students (mean age: 26.95 ± 9.69; 70.02% women). Factor retention methods as Parallel analysis, Exploratory Graph Analysis, and a Categorical Exploratory Factor Analysis revealed a three-dimensional structure of 36-item, as theoretically hypothesized, with desirable internal consistency indices (ω= .868, .906 and .924, respectively). The correlation with measurements of Social Justice and Anti-racism Efficacy suggested validity evidence based on relations with other variables. The results suggest that the instrument is an appropriate measure of CC in sport for adults. Measurement invariance studies with athletes and ethnically diverse populations are suggested.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

380. Functional Social Support assessment in Primary Health Care: a validation study of the DUKE-11-UNC in Spanish Primary health care users.

Irene Gómez Gómez
Universidad Loyola Andalucía/Spain
Juan Ángel Bellón, Isabel Benítez, Patricia Moreno-Peral, Ana Clavería, Joan Llovera, José Ángel Maderuelo-Fernández, Rosa Magallón, Alvaro Sánchez Pérez, Bonabentura Bolibar
Network for Research on Chronicity, Primary Care and Health Promotion (RICAPPS)/Spain
Emma Motrico
Universidad Loyola Andalucía/Spain

Conceptual framework: The Functional Social Support Questionnaire (DUKE-11-UNC) is one of the most used self-reported questionnaires to assess functional social support, especially in primary health care (PHC). Despite the numerous validation studies conducted, there is no consensus about its internal structure. Objectives: The aim of the present study is twofold: 1) to analyse psychometric properties of the DUKE-11 and to obtain validity evidence based on internal structure and on relationships with other variables and 2) to propose scores interpretation. Sample: Two samples of 2997 (S1) and 381 (S2) PHC Spanish users participated in the study. Methodology: Data was collected by face-to-face interviews at different Spanish PHC centres. Exploratory and Confirmatory Factor Analysis (EFA & CFA) were performed to explore the dimensionality. Relationships between DUKE-11 and theoretically related variables (depression, anxiety, and quality of life) were explored through correlation analysis. Reliability was analysed using the Omega coefficient and its 95% CI. Scores were interpreted as percentiles for the total sample and stratified by sex. Results: A one-factor solution was found with EFA and confirmed with CFA (S1CFA; χ^2S-B (44) = 950.549; CFI = .988; NNFI = .985; RMSEA = .070; SRMR = .059; S2; χ^2S-B (44) = 291.548; CFI = .985; NNFI = .9852; RMSEA = .081; SRMR = .072). The one-factor solution presented better-fit indices than the two-factor solutions proposed in previous studies. The theoretical relationships between perceived social support and depression, anxiety and quality of life were found. The 25th, 50th and 75th percentiles corresponded to a score of 40, 47 and 57 points, respectively. In terms of reliability, an adequate Omega coefficient value (ω= %95] ;0.871 CI 878 – 0.864]) was obtained. Implications: The DUKE-11-UNC is a good and valuable tool that can be used to rapidly assess perceived social support in Spanish PHC users.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

285. A psychometric analysis of the Emotion Regulation Questionnaire.

Jennifer Pérez-Sánchez, Ana R. Delgado, Gerardo Prieto
University of Salamanca

The Emotion Regulation Questionnaire (ERQ) is one of the most used tests in the evaluation of the ER. It is a 10-item test that consists of two subscales: cognitive reappraisal (6 items) and expressive suppression (4 items). The present study allowed us to assess the quality of the ERQ scores in a Spanish drivers' sample by means of an invariant measurement model: the Rasch Rating Scale Model (RRSM). A total of 318 male drivers (half of them with road traffic offences and the remaining half, matched controls) participated in this study. Data analysis was carried out using the RRSM. Results indicated that the performance of the response categories was inadequate for both subscales, and so collapsing the 7 original response categories into 3 new categories was necessary. Then the requirement of unidimensionality was met and data-model fit was good enough for both subscales. Item Separation Reliability was over .90, but the parameters of the person's level were estimated with a slightly low degree of accuracy for both reappraisal and suppression.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

309. Transcultural adaptation of the REMICOM Test for the assessment of children's text comprehension monitoring.

Nuria Calet Ruiz
Departamento de Psicología Evolutiva y de la Educación, Universidad de Granada
Jesica Formoso, Juan Pablo Barreyro
CIIPME-CONICET, Argentina
Bárbara Gottheil
INDAGO, Argentina
Gracia Jiménez Fernández
Departamento de Psicología Evolutiva y de la Educación, Universidad de Granada

Comprehensive reading is the ability to decode a text, construct its global meaning, and generate a mental representation of it. It is a complex cognitive skill that requires integrating numerous skills and knowledge, including monitoring comprehension during the reading process. This refers to an individual's ability to regulate their understanding as they read. This study aims to analyze the psychometric properties of a locally adapted version of the REMICOM test, originally designed in Argentina, to assess reading comprehension monitoring in 8- to 10-year-old children. For this purpose, we conducted a pilot study with a sample of 45 third-grade Spanish children from a public school in the city of Granada. We administered a version of the test that included 40 passages, where children were required to identify any problems or contradictory information in them. Out of those statements, 10 were correct, and 30 contained a non-word, an internal inconsistency, or a violation of prior knowledge. We identified the items with the best discrimination capacity to create a final version with 20 items. We obtained measures for validity and reliability. Subsequently, the results were compared with a sample of Argentinean children of the same age. Our findings suggest that the local version of the REMICOM test is a valid and reliable instrument to assess reading comprehension. Additionally, this study contributes to understanding cultural variations in the assessment of reading comprehension monitoring and emphasizes the importance of considering local adaptations in assessment instruments.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

85. Spanish validation of the Dragons of Inaction Psychological Barriers.

Jone Aliri, Laura Vozmediano,
University of the Basque Country UPV/EHU
Laura Pasca
Complutense University of MAdrid
Olatz Goñi-Balentziaga
University of the Basque Country UPV/EHU

Understanding and eventually reducing the gap between environmental concern and effective behavior is key for facing climate challenges. Psychological barriers for adopting pro-environmental behaviors are relevant for understanding this gap, but reliable and valid measures for this construct are not available in many languages. This study aims to adapt and validate to Spanish population the Dragons of Inaction Psychological Barriers (DIPB) instrument. The study used a traditional back translation process, where the first author of the original instrument also participated in comparing the original and the back translated version of the scale. The final Spanish version was administered online to, 337 participants of legal age (18 to 77 years, M = 42.4; SD = 13.6), recruited through several advertisements disseminated through the social media sites of the authors. From the 337 participants, 53.1% (n = 179) identified themselves as men, 46.3% (n = 156) as women, and two as non-binary. The participation in the study was voluntary and anonymous, and approved by the University of the Basque Country's Ethics Committee for Research with Human Beings. One-factor, five-factor and seven-factor structures were compared through nested CFA. The results showed that both the five-factor and seven-factor CFA had a better fit than the single-factor CFA. However, the seven-factor model had a very high RMSEA (greater than 0.08), which questions its use. In addition, model two, the original 5 factor structure, not only has an adequate fit both in terms of CFI and TLI and in terms of RMSEA, but has also adequate standardized factor loadings (>.55), and factor correlations values similar to the original test, ranging from .504 to .754. Although it is necessary to carry out more studies to gather evidence of validity such as gender invariance, and external or predictive validity, this study shows that the Spanish version of the DIPB is a reliable and valid questionnaire.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Construct or concept equivalence

116. The Cross-Cultural Generalizability of Cognitive Ability Measures: A Systematic Literature Review.

Christopher Wilson
Pearson Clinical Assessment / Australia
Stephen Bowden
The University of Melbourne / Australia
Linda Byrne
The Cairnmillar Institute / Australia
Nicole Joshua
Pearson Clinical Assessment / Australia
Wolfgang Marx
Deakin University / AustraliaA
Lawrence Weiss
Test Development Consultant / US

Examining factorial invariance provides the strongest test of the generalizability of psychological constructs across populations and should be investigated prior assessments. The aim of this systematic review was to critically evaluate the current evidence regarding the factorial invariance and the generalizability of cognition models across cultures. The review was structured using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The literature search identified 57 original studies examining the factorial invariance of cognitive ability assessments across cultures. The results were strongly supportive of the cross-cultural generalizability of the underlying cognitive model. Ten studies found configural invariance, 20 studies found weak or partial weak factorial invariance, 12 found strong or partial strong factorial invariance, and 13 found strict factorial invariance. However, the quality of the factorial invariance analyses varied between studies, with some analyses not adopting the hierarchical approach to factorial invariance analysis, leading to ambiguous results. No study that provided interpretable results in terms of the hierarchical approach to factorial invariance found a lack of factorial invariance. Overall, the results of this review suggest that i) the factor analytic models of cognitive abilities generalize across cultures, ii) the use of the hierarchical approach to factorial invariance is likely to find strong or strict factorial invariance, iii) the results are compatible with well-established Cattell-Horn-Carroll constructs being invariant across cultures. Future research into factorial invariance should follow the hierarchical analytic approach so as not to misestimate factorial invariance. Studies should also use the Cattell-Horn-Carroll taxonomy to systematize intelligence research.

Columbia (Canada) ***Shared, Insights and Unique Perspectives: What Can Psychological Testing and Assessment Learn from Innovations in Survey Research and Vice Versa***
Poster Session 3
Topic: Construct or concept equivalence

146. Measurement Invariance of the Job Insecurity Scale Across Gender in Aging Workers.

Lei Liu
Renmin University of China
Yuan Jiang
Beijing Sport University
Jianmin Sun
Renmin University of China
Ping Fang
Capital Normal University

As aging intensifies, the proportion of aging workers in organizations continues to rise, highlighting psychological health issues, especially job insecurity. Job insecurity refers to an individual's apprehension about the potential loss or deterioration of their current job or key job characteristics, such as position and compensation. Research indicates that job insecurity not only engenders negative attitudes toward one's job and organization but also impairs the physical and mental health of employees. Understanding and addressing the job insecurity experienced by aging workers is crucial for promoting successful aging at work. Therefore, assessing job insecurity among aging workers of different genders is necessary. The Job Insecurity Scale (JIS), which includes dimensions of quantitative and qualitative job insecurity, is widely used. However, there has been no study to date examining its gender invariance among aging workers. Since gender invariance is a prerequisite for comparing scores between genders, testing the gender invariance of the JIS is essential. This study aims to examine the gender invariance of the JIS among aging workers and to compare gender differences. The study surveyed 1,135 aging workers using the JIS and conducted both single-group and multi-group confirmatory factor analysis. The results showed that the models of configural, metric, scalar, and strict invariance of the scale were established across different genders of aging workers. Latent mean comparisons revealed no significant difference in quantitative job insecurity scores between female and male aging workers, while qualitative job insecurity scores were lower for female aging workers. This study demonstrates that the JIS possesses cross-gender measurement invariance among aging workers. Research on gender differences in job insecurity among aging workers in the post-pandemic era holds significant practical implications for organizations.

WEDNESDAY 3 JULY
Session 6.1 SYMPOSIUM
Topic: Innovations in test development

186. Continuous Norming: Recent Advancements in Research and Application – Part B

Jan-Philipp Freudenstein
Hogrefe Publishing Group

Continuous norming refers to various methods of statistically modeling psychological test scores in relation to predictor variables (e.g., age) to produce standardized test scores. It is an increasingly popular approach for increasing the precision of norm scores by avoiding artificial categorization of predictors. In recent years, research has provided several methodological advances in continuous norming that have greatly increased the utility of these procedures for test developers. Despite its success, several challenges and potential improvements remain in the methodology and application of continuous norming. This two-part symposium will address these remaining issues by providing insights into ongoing research and practical applications. The symposium thereby aims to provide a comprehensive understanding of continuous norming as well as in-depth insights into current research and applications. In particular, this second part of the symposium will provide recent methodological research on continuous norming. This includes a method to adjust for non-representativeness of samples, optimal sample size calculations, item response theory based norming, and research on more efficient continuous norming by using prior information. Overall, the symposium may guide future research, improve test development practices, and set new standards for reporting in the field of psychological testing.

WEDNESDAY 3 JULY
Session 6.1 SYMPOSIUM
Topic: Innovations in test development

196. Sample size calculation and optimal design for univariate and multivariate regression-based norming (Innovations in test development)

Francesco Innocenti, Math Candel, Frans Tan, Gerard van Breukelen
Maastricht University

To classify subjects' performance on relevant clinical or educational measures, such as neuropsychological tests or language tests, psychologists and educators need reference values or norms. Norms are important because they allow practitioners to compare subjects' performance on a test with the reference population determined by the purpose of the test (e.g., diagnosis of dementia versus job selection). Based on this comparison, psychologists and educators can make important decisions about individuals, such as assigning a patient to a treatment. For this reason, norms must be precise, that is, not strongly affected by sampling error in the sample on which the norms are based. In this talk, it will be shown how to minimize sampling error and maximize the precision of the norms in three steps. First, norms should be obtained by regressing the test score on predictors deemed relevant for the norming (e.g., age and sex), since this approach is more efficient than the traditional approach of splitting the sample into subgroups based on demographic factors and deriving norms per subgroup. Second, the precision of the norms can be maximized by carefully selecting the composition of the sample of the normative study (e.g., which age groups to include) for a given total sample size. Third, this sample size can be computed such that a prespecified power for classification, or a prespecified precision of estimation, is obtained. These three steps will be illustrated under two scenarios: norming of one test with multiple linear regression and norming of several tests with multivariate multiple linear regression.

WEDNESDAY 3 JULY
Session 6.1 SYMPOSIUM
Topic: Innovations in test development

265. Item Response Theory Based Continuous Test Norming (Psychometric modeling)

Hannah Heister, Casper Albers
University of Groningen/Netherlands
Marie Wiberg
Umeå University/Sweden
Marieke Timmerman
University of Groningen/Netherlands

Even though item response theory (IRT) based models are known to result in better latent trait estimates than raw scores (e.g., sum score) do, current continuous norming methods rely on raw test scores. To profit from the information in individual item scores, we propose 2PL-norm: a Bayesian two-parameter logistic IRT model with age-dependent mean and variance of the latent trait distribution. The norms are then derived using the estimated latent trait score and the age-dependent distribution parameters. In a simulation study, we evaluated the performance of 2PL-norm and compared it with the currently leading continuous norming methods cNORM and GAMLSS. The simulation shows that 2PL-norm has a better overall performance than cNORM and GAMLSS, and that 2PL-norm is less effective at the tails of the latent trait within each reference population. Moreover, the credible intervals from 2PL-norm, expressing error due to measurement and sampling variability, have a much better coverage than the confidence intervals of cNORM and GAMLSS. An application of 2PL-norm to normative data from an intelligence test demonstrates that the model can fit empirical data and provides greater insight into the latent trait changes over age compared to the current continuous norming methods. For empirical practice our study suggests that test constructors solely interested in the extreme trait positions should use GAMLSS-based norming while those interested in the full reference population should use 2PL-norm.

WEDNESDAY 3 JULY
Session 6.1 SYMPOSIUM
Topic: Innovations in test development

221. Adjusting for non-representativeness in continuous norming with Multilevel Regression and Poststratification (Psychometric modeling)

Klazien de Vries, Marieke Timmerman, Anja Ernst, Casper Albers
University of Groningen

In psychological test norming, non-representativeness of influential background variables in the normative sample leads to bias in the normed score estimates. Representativeness is difficult to establish in practice, hence adjustment methods are needed to combat this bias. We combined multilevel regression and poststratification (MRP) with regression based norming to derive continuous norms from non-representative normative samples. This approach is then compared to the current adjustment methods in continuous norming: weighted regression and weighted cNORM; in addition, regression with post-stratification (RP), a counterpart of MRP, is considered. The results of our simulation study indicate that MRP is more efficient than weighted regression and weighted cNORM and is slightly more robust than RP. We illustrate the use of MRP using empirical scores from an intelligence test. We argue that MRP is a valid adjustment method in regression based norming, and recommend its use to mitigate bias in non-representative normative samples.

WEDNESDAY 3 JULY
Session 6.1 SYMPOSIUM
Topic: Innovations in test development

384. More efficient continuous test norming by using prior norm information (Psychometric modeling)

Lieke Voncken
Tilburg University
Thomas Kneib
University of Göttingen
Casper Albers
University of Groningen
Nikolaus Umlauf
University of Innsbruck
Marieke Timmerman
University of Groningen

In continuous norming, (psychological) test scores are typically estimated in relation to predictor variables (e.g., age). To estimate this relationship properly, large normative samples may be needed. In this talk, we will discuss to what extent this burden can be alleviated by using prior information in the estimation of new norms. In a simulation study, we investigated using Bayesian Gaussian distributional regression to what extent this norm estimation is more efficient and how robust it is to prior model deviations. We varied the prior type, prior misspecification and sample size. In the simulated conditions, using a fixed effects prior resulted in more efficient norm estimation than a weakly informative prior as long as the prior misspecification was not age dependent. With the proposed method and reasonable prior information, the same norm precision can be achieved with a smaller normative sample, at least in empirical problems similar to the simulated conditions. This may help test developers to achieve cost-efficient high-quality norms. We illustrate the method using empirical normative data from the IDS-2 intelligence test, and we discuss the general implications of this proof of concept for using prior norm information.

WEDNESDAY 3 JULY
Session 6.2 SYMPOSIUM
Topic: International assessment

595. Debating Foundational Competencies in Educational Measurement: International Perspectives on an NCME Task Force Consensus

Andrew Ho
Harvard University

What are "foundational competencies in educational measurement"? What knowledge, skills, and abilities must current students of educational measurement possess in order to succeed and continue learning in our field? In 2023, a published article from a Presidential Task Force of the National Council on Measurement in Education (NCME) attempted to answer these questions. The Task Force identified three competency domains and five competency subdomains. The article also demonstrated how educational measurement careers and curricula develop these competencies. In this symposium, ITC members will put the Task Force Report to the test, challenging the framework and its elements, and offering international perspectives on foundational competencies that educational measurement requires. Andrew Ho, who chaired the Task Force, will lead off the presentations by describing the Task Force consensus framework. Three ITC members, Isabel Benitez, Dragos Iliescu, and Lisa Keller, will then present their own commentaries on the framework, answering three common questions: 1) Is the framework coherent, defensible, and useful? 2) Are any foundational competencies missing, superfluous, or unambitious? 3) How can or should the International Test Commission, its members, and its constituencies continue to support consensus around foundational competencies in educational measurement? Alina von Davier and Howard Everson, Task Force members, will respond to the presentations, engage the full panel of presenters in brief debate, and then solicit questions and comments from the audience. By encouraging debate and possible consensus among symposium members about the foundations of our field across countries and cultures, this symposium has a goal aligned with the conference theme: Working together to improve cross-cultural assessment and research.

Discussant name: Alina
Discussant surname: von Davier
Discussant affiliation: Duolingo

WEDNESDAY 3 JULY
Session 6.2 SYMPOSIUM
Topic: International assessment

596. Navigating the Foundational Competencies in Educational Measurement: Enhancing validity, validation, and fairness through comprehensive approaches (International Assessment)

Isabel Benitez
University of Granada

The consequences derived from educational measurement exert a direct and profound impact on individuals' lives, shaping the trajectory of their academic and professional journeys. It is imperative to ensure that professionals in educational measurement possess the necessary competencies to effectively address the challenges inherent in their work. The National Council on Measurement in Education Presidential Task Force has identified three foundational competencies crucial for educational measurement programs and professions. This presentation focuses on the "Validity, Validation, and Fairness" subdomain (Subdomain B) within the broader Educational Measurement Competences outlined by the National Council on Measurement in Education Presidential Task Force. The discussion centers on the importance of incorporating comprehensive approaches when addressing validity and designing validation studies. Two primary ideas will be explored: a) the alternative and complementary contributions of validity evidence based on response processes and testing consequences; and b) the role of qualitative evidence gathered through qualitative or mixed designs to enhance understanding of the measure. These considerations aim to elucidate potential interconnections between Subdomain B and the subdomain "Context: Social, historical and political." Additionally, the presentation introduces a previously overlooked competence: the ability to identify measurement needs and utilize them to guide decision-making during validation processes. Furthermore, the presentation examines the advantages and disadvantages of the proposed framework within the educational system context, specifically in Spain. It concludes with proposals for creating a teaching framework that promotes meaningful learning through exposure to both real and simulated work situations.

WEDNESDAY 3 JULY
Session 6.2 SYMPOSIUM
Topic: International assessment

597. On the future integration of foundational competencies frameworks in connected professions (International assessment)

Dragos Iliescu
University of Bucharest

The "Foundational Competencies in Educational Measurement" FCEM project is an important initiative that is part of a growing body of literature looking into the competencies associated with professions in the domain of social and behavioral sciences. Here, the focus is on educational measurement and the set of foundational competencies it is composed of. We salute the publication of this dedicated framework. At the same time, we signal a number of points that could further improve this initiative. We especially underscore two points. The first point is a matter of integration/relation to existing frameworks we will especially discuss possible integration with the IPCP Declaration International Declaration of Core Competences in Professional Psychology as a generalist professional competencies framework and the EFPA EuroPsy as a measurement-focused competencies framework. These both are important initiatives that already function and could be very well integrated with the FCEM, to the benefit of an international audience. The second point relates to the structure of the FCEM, where we draw attention to two aspects that could be addressed to clarify the structure of the FCEM and maybe also sharpen its applicability: a possible overlap between domains and subdomains within the competence model; and b independence and delineation from previous discussions and extant/established research about the structure of professional competencies, such as the famous "competency cube" proposed by Rodolfa 2005.

WEDNESDAY 3 JULY
Session 6.2 SYMPOSIUM
Topic: International assessment

598. What is truly foundational: Continuing the conversation on the Foundational Competencies in Educational Measurement (International assessment)

Lisa Keller
University of Massachusetts Amherst

Developing a framework regarding the core competencies in educational measurement is an important and very difficult task. The Task Force has produced a starting point for that very important conversation. It is challenging to develop such a framework that meets the needs of all stakeholder groups and reflects the practices and values of all training programs for measurement professionals. We applaud the task force for undertaking this endeavor and starting the conversation. Our review and critique of the framework considers the opinions and perspectives of both faculty and graduate students. In our response to the framework, we offer several aspects to be considered in the revision of the program. Fundamentally, there appears to be a need to create more comprehensive definitions of the terms used in the framework as they appear narrowly defined to reflect past and current practice rather than ideal future-looking practice. Along with that, a better understanding of what is meant by "competency" is needed; what is the depth of knowledge that signals competency? In addition to the need for a clearer vision of the document, we felt there were other areas that needed addressing including (1) how to be sure the content is forward looking and not a reification of the past, (2) the role of international perspectives, (3) the role of AI and machine learning, and (4) the role of psychometricians in the global educational crisis of basic skills and literacy. We look forward to being part of the conversation as our field undertakes this task.

WEDNESDAY 3 JULY
Session 6.2 SYMPOSIUM
Topic: International assessment

599. **Foundational Competencies in Educational**

Andrew Ho
Harvard University

This article presents the consensus of a National Council on Measurement in Education Presidential Task Force on Foundational Competencies in Educational Measurement. Foundational competencies are those that support future development of additional professional and disciplinary competencies. The authors reviewed job postings, course syllabi, and classroom activities to develop a framework for foundational competencies in educational measurement. The article illustrates how educational measurement programs can help learners develop these competencies. It also describes how foundational competencies continue to develop in educational measurement professions. Word limits prevent a full review of the Task Force consensus, but a written description follows: The framework introduces three foundational competency domains: 1) Communication and Collaboration Competencies; 2) Technical, Statistical, and Computational Competencies; and 3) Educational Measurement Competencies. Within the Educational Measurement Competency domain, the authors identify five subdomains: 3A) Social, Cultural, Historical, and Political Context; 3B) Validity, Validation, and Fairness; 3C) Theory and Instrumentation; 3D) Precision and Generalization; and 3E) Psychometric Modeling. The full article is available here and includes a glossary and illustrated framework in Figures 1 and 2, respectively: https://onlinelibrary.wiley.com/doi/epdf/10.1111/emip.12581 The authors make the case that foundational competencies are not exhaustive, that they overlap and intersect with each other, and that they should be both descriptive of the field and aspirational about the future. The article presents an illustrative first-year course in educational measurement and provides examples of how careers in educational measurement both require and continue to develop these competencies.

WEDNESDAY 3 JULY
Session 6.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

723. Exploring Human Dimensions - Love, Moral Dilemmas, Well-being, and Cultural Sympathy

Alejandra Dominguez Espinosa
Universidad Iberoamericana

The aim of the symposium is to unraveling the complexities of our emotions, ethical considerations, mental well-being, and the rich tapestry of cultural influences that shape our social interactions. Our symposium will delve into four pivotal topics, each shedding light on distinct aspects of the human psyche. We will be traversing through the realms of emotion, morality, mental wellness, and cultural nuances, providing a holistic understanding of the intricate web that weaves together our shared human experience. Firstly, we will explore the profound depths of human emotion through the lens of an "Adapted Love Scale." Love, a universal force that transcends borders and cultures, will be examined in its nuanced forms and expressions. Our distinguished speaker will guide us through the adaptation of this scale, emphasizing its cultural relevance and potential implications for mental and emotional well-being. Following this, our symposium will navigate the moral landscapes that define our ethical compass. A "Developed Scale of Moral Dilemmas" will be presented, unraveling the complexities of ethical decision-making. This exploration promises to offer insights into the intricate interplay between personal values, societal norms, and the moral fabric that binds us together. The symposium will then transition to an analytical perspective, delving into the "Psychometric Analysis of the Well-being Survey." This critical examination seeks to decipher the intricacies of our mental well-being, utilizing rigorous psychometric methods to unveil patterns and insights that can inform interventions and policies for a healthier society. Lastly, we will focus on the "Cultural Development of the Sympathy and Agreeableness Scale." As we traverse through diverse cultural landscapes, we will explore how perceptions of sympathy and agreeableness are shaped, providing a profound understanding of the cultural underpinnings that influence our social interactions.

WEDNESDAY 3 JULY
Session 6.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

770. Validation of the Romantic Love Myths Scale of Bonilla-Algovia and Rivas-Rivero in a group of Mexican participants (Translation of tests, psychological assessment instruments and survey questionnaire)

Carolina Armenta Hurtarte
Universidad Iberoamericana
María Bárbara Rivero Puente
Technolochicas

Love, as a complex emotion, manifests in various forms such as brotherly love, parental love, self-love, and erotic love. This study focuses on romantic love and its potential evolution into violent dynamics within intimate partner relationships. Addressing intimate partner violence (IPV) is crucial, necessitating a deeper understanding of its causes. This paper posits that exploring romantic love provides a valuable lens for comprehending IPV, given the links between the idealization of romantic love, the romantic love model, and the perpetuation of patriarchal social structures and gender inequalities. The primary objective of this research was to validate the Bonilla-Algovia and Rivas-Rivero romantic love scale in a Mexican population. The non-random intentional sampling involved 375 participants aged 18 to 40 years. Subsequently, the participants were divided into two subsamples: Group 1 comprised 196 individuals aged 19 to 55 years (Mean=25, SD=6.03), while Group 2 included 179 participants aged 18 to 52 years (Mean=23, SD=4.69). The reliability of the total scale, consisting of 8 items, was assessed using Cronbach's alpha (α = .780), indicating good reliability. Factor analysis revealed two factors: (F1) Idealization of love with α = .738 and (F2) Abuse/love with α = .738, both demonstrating satisfactory internal consistency. Despite the valuable insights provided by this study, certain limitations should be acknowledged. Geographical data for participants beyond their Mexican nationality was not collected, limiting the precision of the study's scope. Future research should aim to enhance the scale's utility by exploring additional elements and considerations. In conclusion, this study contributes to the understanding of the complex interplay between romantic love and intimate partner violence. The validated Bonilla-Algovia and Rivas-Rivero romantic love scale offers a valuable tool for researchers and practitioners seeking to delve deeper into relationships.

WEDNESDAY 3 JULY
Session 6.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

735. Assessing tolerance to corruption through moral dilemmas: psychometric properties of a measurement (Translation of tests, psychological assessment instruments and survey questionnaire)

Alejandra del Carmen Domínguez-Espinosa
Universidad Iberoamericana Ciudad de México
Christian Cruz-Torres, Tonatiuh García-Campos
Guanajuato University
Irene Salas-Menotti
Fundación Universitaria Konrad Lorenz
Carlos Alberto Montesinos-González, Pablo Domínguez-Perera
Universidad Iberoamericana Ciudad de México

Corruption has long been a pervasive issue in the historical context of Mexico. Understanding the extent of citizen participation in corrupt practices presents a formidable challenge due to the illicit and socially reprehensible nature of corrupt acts, which inherently introduces a social desirability bias into attempts to measure corruption (Krumpal, 2013). Individuals are naturally inclined to downplay their involvement in such activities and express a willingness to combat them. On the other hand, acts of corruption involve the possibility of cooperating with other people to obtain common benefits, so refusing to participate often implies high social costs (Muthukrishna et al., 2017). Moreover, corruption is not a uniform experience; it varies significantly among individuals based on their unique opportunities to engage in corrupt behavior. The goal of this study was to design a measurement tool to assess individuals' propensity to either tolerate or resist corruption, employing moral dilemmas (Christensen & Gomila, 2012) that include the cooperative nature of corruption acts. The instrument was administered to 173 participants for the exploratory factor analysis, followed by 282 participants for the confirmatory factor analysis. The results of exploratory factor analysis revealed a single factor that grouped seven out of the eight dilemmas, obtaining satisfactory levels of reliability (χ=.71) and goodness-of-fit indices (Chi2=20.69, p=.11, CFI=.95, GFI=.99, TLI=.94; RMSEA=.08, 90% CI [.03, .11]). This factor structure was subsequently validated through a confirmatory factor analysis, reinforcing the instrument's robustness and validity (Chi2=10.58, df=14, p=.71, CFI=1, GFI=.99, SRMR=.03, RMSEA< .001, IC 90% [.001, .04.]. It is concluded that this instrument is useful for future investigations across contexts where similar characteristics of corruption tolerance and resistance need to be measured.

WEDNESDAY 3 JULY
Session 6.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

731. The measurement of subjective well-being in Mexico: A psychometric analysis of the ENBIARE 2021 (Validity theory in testing, psychological assessment and survey research)

José Luis López Silva
UNAM

Subjective well-being (SWB) is a psychological construct that refers to the various types of subjective evaluations of one's life, including both cognitive evaluations and affective feelings (Diener, Lucas, et al., 2018). Its systematic measurement provides important information about the quality of life in societies and several international institutes and organizations such as the OECD (2013) have promoted the implementation of these measures in large-scale surveys, the ENBIARE being the result of these efforts in Mexico. The present work consisted of a psychometric analysis of the SWB section of said survey. A cross-sectional survey research study was carried out using INEGI databases, assessing whether it met the adequate psychometric standards for the measurement of the construct. Evidence of content validity and internal structure was evaluated, along with evidence of precision and fairness. It was observed that this section has relevant and representative items for the most part and that they are supported by international guidelines. Exploratory and confirmatory factor analyses showed configurations congruent with theory and research in the area, specifically within the hedonic conception of well-being and Diener's (1984) tripartite structure; with the structural equation models, however, in the framework of this study it was not possible to empirically differentiate between top-down and bottom-up theories of SWB and different models are required to contribute to the debate about the causal processes that produce the structure of cognitive well-being (Schimmack, 2008). Favorable reliability coefficients and evidence of invariance for some sociodemographic characteristics were also found. Finally, the relevance of using the ENBIARE for public policy and psychological research was discussed. Keywords: subjective well-being, psychometrics, secondary data, validity, public policy

WEDNESDAY 3 JULY
Session 6.4 SYMPOSIUM
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

799. Proposal for Ethnopsychological Measurement to Estimate Agreeableness and Sympathy in Mexico (Validity and fairness in cross-cultural testing, psychological assessment and survey research)

Vanessa Edith Arellano Carranza
Universidad Nacional Autónoma de México

For the development of new theoretical and methodical models of personality, Church (2000) proposes a joint integration of trait psychology and culture; being the Five Factor Model (McCrae & John, 1990) one of the most centralized measurement instruments for the study of personality where the dimension of Agreeableness (etic) has generated evidence of meaning in the adaptations of Latin social context, a fact that has influenced the study of the culturally relevant dimensions for the understanding of the relationship in this trait (emic) as the study of Agreeableness in the context has been presented The main axis of this study is a proposal for ethnopsychological estimation of Agreeableness and Sympathy as personality traits in Mexico with an emic-etic study perspective (Cheung et al. 2011 and Díaz Guerrero, 1994). By means of qualitative exploration for the design of psychometric scales for the estimation of these traits in our Mexican of 294 participants with the collection of validity evidence referred to the content (judging by experts in social psychology), as well as evidence of precision validation (reliability analysis with Cronbach's Alpha, Ordinal Alpha, Hierarchical Omega and Total Omega), validity focused on the internal structure (Exploratory Factor Analysis) and validity evidences referring to the relationship with other variables (nomological network) with the styles of Coping and Stress Management (Correlation Analysis and Multiple Linear Regression). The main contributions of the study is the development of two psychometric scales, Agreeableness with factors I. Altruism and II. Agreeableness, Sympathy with Factors I. Consideration for Others and II. Cordiality, as well as the estimation of the reliability factors that comprise the ordinal nature of the data from the application of the RStudio development environment (versions 4.0.2 and 4.1.1), being the evidences of positive influence of Agreeableness and Sympathy in the style of Confrontation I.

WEDNESDAY 3 JULY
Session 6.5
Topic: Innovations in test development

546. **Advancing Precision and Validity with CAT and NLP**

Alexandre Jaloto
National Institute for Educational Studies and Research Anísio Teixeira (Inep)

This work consolidates three studies demonstrating advancements in precision and validity evidence for educational tests, employing data from Brazil's National High School Examination (Enem). The Enem consists of tests of four knowledge area with multiple-choice items and is used in Brazil for university admissions. In the initial study, we simulated the application of Computerized Adaptive Testing, comparing its precision with the traditional 2020 Enem and observing a consistent reduction in the standard error of measurement across all ability levels and tests, with a reduction of more than 50% on the test length. The second study focused on the interdisciplinary nature of Enem content. Using Natural Language Processing (NLP) tools, we characterized interdisciplinarity by analyzing item text similarity through an exploratory graphical analysis of word vectors employing the Glove algorithm with 300 dimensions. Each item was positioned on a two-dimensional graph, revealing the distinct identities of each Enem area and their respective curricular components based on item proximity. Additionally, analysis showed interdisciplinary connections between items of different knowledge areas. The third study extended our exploration to Natural Science test items from 2009 to 2020, employing topic modeling configured for three topics. These topics were notably characteristic of the disciplines of Biology, Physics, and Chemistry, and items highly characterized by more than one topic presented interdisciplinary content. The findings of the second and third studies contribute significantly to validity evidence based on the content of Enem tests as they showcased interdisciplinarity and the approach to specific curricular disciplinary topics. Our intent is to engage in a comprehensive discussion on the potential of CAT and NLP tools to significantly enhance evidence of precision and validity in the realm of educational testing.

WEDNESDAY 3 JULY
Session 6.5
Topic: Innovations in test development

691. Privacy and Access to Data in Chile's National Admission Test: Balancing Students´ Rights and Institutional Needs

Agustin Barroilhet, Francisco Lechuga
University of Chile
Monica Silva
Catholic University of Chile
Leonor Varas
University of Chile

Access to educational admissions data is essential for stakeholders such as test developers and university administrators. System-wide analyses are imperative for evaluating the efficacy of entrance exams in forecasting college success. Such analyses depend on the sharing of sensitive applicant data, including test scores, academic performance, demographics, and financial backgrounds. In Chile, this necessitates a collaborative effort among testing agencies, university authorities, and the Ministry of Education, which has governed the admission system's integrity since 2020. However, conflicting incentives among these bodies create barriers. The Ministry is wary of the entrance exam being misused to unfavorably rank public high schools against their private counterparts. Non-selective universities resist collaboration, fearing that a fortified entrance exam could disadvantage them in attracting top students. Consequently, the onus to enhance the exam rests with the testing agency and selective universities. This complex landscape has rendered predictive studies—once facilitated by lax data privacy practices—increasingly difficult, even with anonymized data. This paper proposes a strategy to navigate the complexities surrounding the use of admission test data, addressing potential misuses and advocating for legal reforms that clarify exemptions in privacy legislation. It argues for the admissions test's role as a public good, asserting that its benefits should not be overshadowed by short-term interests or misinterpretations of privacy statutes. The lessons offered might be useful for other countries facing similar dilemmas.

WEDNESDAY 3 JULY
Session 6.5
Topic: Innovations in test development

718. Measuring Communication and Interpersonal Skills of Pre-Service Teachers: An Experimental Study Using Frontal Alpha Asymmetry Indicator

Ahmet Haphap, Nilüfer Kahraman Gazi
University/Turkey

The ability to communicate effectively is an essential skill for teachers' professional development. Recent years have brought a growing awareness of the necessity of training programs taking an active role in supporting pre-service teachers' Communication and Interpersonal Skills (CIS). Accurately assessing CIS components, however, such as empathy and relational versatility has long been a real challenge. Researchers have been working to create innovative forms of assessments that would be relatively free from response biases. Within this context, our study presents a scenario-based scale format, which was created as a part of a larger research project exploring various alternative assessment tools for measuring CIS. The format makes use of a developing storyline where the participant assumes a teacher's role within a two-part narrative and responds to two multiple-choice items. In part 1, the story starts and then the participant is asked to choose a response (Item 1, measuring empathy). In Part 2, the story continues with the character's reaction to the first response. Then the participant is asked to choose a response (Item 2, measuring versatility). Study data were collected from 43 pre-service teachers in a lab setting, enhanced with the Electroencephalogram (NE Enobio 8 EEG headset). Results show that the correlations between mean Frontal Alpha Asymmetry (FAA) indicators and the item responses were higher for Item 2 (versatility) when compared to those of Item 1 (empathy). This is consistent with the expectation that the FAA correlations would be higher for the versatility item when compared to the empathy item, for the former requiring a more complex emotional processing skill than the latter. Albeit limited due to the experimental sample size of the study, our findings warrant a further study of the FAA correlates of items used in new assessment formats, especially when they are meant to measure subskills composing a complex latent trait, such as CIS. This study was partially supported by Gazi School of Education and by TUBITAK under grant SOBAG 120K142.

WEDNESDAY 3 JULY
Session 6.5
Topic: Innovations in test development

780. Taking an Organimetric approach to Organisational Change

Nigel Evans
NEC, UK
John Mervyn-Smith
GC Index, UK

Aims: Many aspects of traditional psychometric testing do not fully resonate with OD practitioners and their business clients, who are more familiar with Organisational Survey methodology. This paper presents the development of an Organimetric Index that measures and describes energy for impact through proclivity, rather than personality. The language and concept of the Organimetric approach is shown to be practical, as it applies at the individual, team and organisational level. Objective: Organisational growth requires effective transformation, particularly identifying people who think differently. Responding to a funded corporate client request, researchers were able to scientifically investigate what makes a 'game changer' from organisational field study. Methods: Research followed 3 main phases over three years: Phase 1 was an initial exploration of the characteristics of 'Game Changers' using Repertory Grid interviews which yielded 180 observations. Phase 2 built on phase 1, suggesting defining characteristics for Game Changers under two broad constructs Imagination and Obsession. Phase 3 expanded into Game Changer self-perception data. Factor analysis of 1000 questions was completed to explore the meaning of different responses to the questionnaire. Results: The research yielded reliable and valid measures of Game Changers alongside four additional distinct proclivities when it came to making an impact at work. These differences are now represented as an Organimetric Index identifying both individual and collective impact of how people contribute to achieving transformation. International sampling has found very few differences in the proclivity scores, more data is being gathered to confirm. Implications: Application case studies show the Index provides a clear framework that can be used to inform key business decisions in an inclusive way by acknowledging the diversity of thought and action that come from different proclivities.

WEDNESDAY 3 JULY
Session 6.5
Topic: Innovations in test development

789. Qualitative procedures in developing a new emotion regulation measure for children from 4 to 6 years old

Denise Ruschel Bandeira, Aline Riboli Marasca, Gabriela Prestes, Gabriela Nunes Maia
Universidade Federal do Rio Grande do Sul

The ability to regulate emotions effectively is one of the goals for a child's socioemotional development. Assessing emotional regulation is a way to monitor the child's development and propose early interventions. There are a substantial number of instruments aimed at assessing children's emotional regulation, with the majority relying on adult reports. However, the existence of instruments that enable evaluation from the child's perspective is also necessary. To address this gap, a new tool for assessing emotion regulation in preschoolers is currently in progress. The present study aims to delineate the initial qualitative procedures involved in developing this instrument. First, a qualitative study was undertaken with 12 Brazilian children aged 4 to 6 years (58.3% girls). In a semi-structured interview, children were asked to describe situations in which they felt anger, sadness, and fear. They also had to indicate what they would think or do to stop the emotion depicted in three vignettes. Three evaluators independently performed the content analysis, and categories were derived from prior studies. The results indicated that social situations were the most common elicitors of emotions. Children employed strategies such as problem-solving, distraction, self-comfort, and seeking help more frequently. A novel emotional regulation strategy termed 'magic solution' emerged. In the subsequent phase, emotional eliciting vignettes were crafted based on children's responses. The most frequently cited emotion regulation strategies were categorized, with definitions and examples provided, forming the basis for analyzing responses to the vignettes. Both the vignettes and response categories will undergo expert analysis to ascertain content validity. These procedures form an empirical base for instrument development, aligning with the literature review. They aid in adapting language for preschoolers, including situations familiar to the target audience.

WEDNESDAY 3 JULY
Session 6.6
Topic: Translation of tests, psychological assessment instruments and survey questionnaire/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

229. The association between changed translations and item functionality in ICCS 2022

Anna Wikstrom, Lauren Musu
IEA

Research has established that question wording can influence how individuals respond to an assessment item, yet little is known about the impact of changing trend item translations between cycles in international large-scale assessments (ILSAs). Understanding the impact of translation changes can bring valuable insight into the validity and reliability of ILSAs. This exploratory study used data from 9 countries which participated in the 2016 and 2022 International Civic and Citizenship Education Study (ICCS) to investigate whether changes to translations between cycles are associated with changes in item functionality measured by differential item functioning (DIF). ICCS provides insights into how school systems across the world help prepare adolescents to undertake their role as citizens. It contributes to the knowledge of how young people understand a variety of topics, such as environmental sustainability, global and digital citizenship, and migration. Data for the current study were comprised of 55 trend items which aim to measure change over time and thus were identical in both cycles. Preliminary results suggest that the types of translation changes observed for ICCS did not have a statistically significant impact on item functionality at the international level, possibly due to considerable in-group variation. However, the range in DIF was larger for changed translations indicating that such changes could be impacting item functionality and therefore should be implemented with caution. The results are discussed in light of methodological challenges and the potential cross-level interaction between time and translation changes. In addition, external factors such as the covid-19 pandemic and political unrest in the West could have impacted the functioning of trend items. Together, these factors contribute to understanding response behavior in ILSAs across time.

WEDNESDAY 3 JULY
Session 6.6
Topic: Translation of tests, psychological assessment instruments and survey questionnaire/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

481. Going Global: 39 language versions of the BFI-2-XS

Beatrice Rammstedt, Matthias Bluemke, Clemens Lechner, Lena Roemer, Dorothée Behr
Deutschland
Steve Dept, Laura Wayrynen
cApStAn
Christopher Soto
Colby College
Oliver John
University of California at Berkeley

Across recent decades, the Big Five have increasingly been accepted as a framework to parsimoniously describe personality on a global level (e.g., John et al., 2008; McCrae & Costa, 2008). This has led to broad interest in their assessment, even in fields beyond core personality research, such as sociology, economics, and epidemiology. Also, numerous cross-cultural surveys now include measures for the broad personality dimensions in order to compare cultural differences in their impact on the constructs of interest in these studies. In the 2023 Survey of Adult Skills (PIAAC), the Big Five personality traits were assessed via the15-item extra-short version of the Big-Five-Inventory-2 (BFI-2-XS; Soto & John, 2017). For this purpose, the instrument was translated and adapted into 25 languages and to 29 countries resulting in 39language versions. This translation and adaptation process followed state-of-the-art procedures to generate language versions of the BFI-2-XS that are maximally comparable across countries and regions. In the presentation, we will describe this general translation procedure from a methodological point of view. We also document each resulting language version and report in detail which decisions were taken during the translation process and which adaptations were made to existing national versions of the BFI-2-XS. By that, we aim to share the resulting high quality and cross-culturally comparable national BFI-2-XS versions with researchers and enable their reusage.

WEDNESDAY 3 JULY
Session 6.6
Topic: Translation of tests, psychological assessment instruments and survey questionnaire/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

612. Exploring Careless Response Behaviours in Surveys: A Comparison of Different Identification Methods

Murat Doğan ŞAHİN, Basak Erdem Kara
Anadolu University/Turkey

Careless responding poses a major threat to self-report measures' validity, as evidenced by the recent surge in relevant studies. However, it can be stated that identifying careless response behavior is quite complex. Since there are various manifestations of careless responding, a number of descriptive indices have been proposed. Some of these are based on direct measures (e.g., instructed items) or on indirect measures such as response time, longstring, intra-individual response variability and response consistency (Goldhammer et al., 2020). Alternatively, model-based indicators offer a holistic approach to assessing response quality. In this regard, IRT-based methods, especially person-fit statistics, have been employed to detect careless responses (Niessen et al., 2016). However, when IRT's stringent assumptions are unfeasible, non-parametric methods offer benefits, as demonstrated in Wind and Wang (2022) and Wind et al. (2023)'s literature-enriching studies. Within the scope of this study, the aim is to compare the performance of certain direct and indirect measures with model-based indicators in identifying careless respondents. Accordingly, it is intended to use data obtained from a multi-dimensional self-report measure, which is reliable, valid and has a very wide field of use, to approximately 400 individuals. In this context, from the descriptive approaches, instructed item, longstring and intra-individual response variability will be compared with model based approaches, that is traditional IRT person-fit statistics and the Mokken analysis-based approaches proposed by Wind et al. (2023). Additionally, the sequential procedure proposed by the same researchers will be applied, that is based on detecting and removing careless respondents at each step. This procedure will be implemented until the number of flagged careless respondents by these statistics will decrease sufficently.

WEDNESDAY 3 JULY
Session 6.6
Topic: Translation of tests, psychological assessment instruments and survey questionnaire/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

677. SWIFT: Developing a Digital Psychometric Test for Swift Identification of Students at Risk for Learning Difficulties and Underachievement in Schools

Rebecca Good, Dr. Kate James, Trevor James
Education Elephant Ltd

Approximately 60% of students with an attention and/or learning difficulty pass through school unidentified. The consequences can range from underachievement, missed opportunities for support, behavioural problems, withdrawal, lack of motivation and lowered mental health. While there are many tests available to schools many are too labour intensive, time consuming or lack the necessary data to identify these children. The authors currently produce a group administered, pen and paper test, the SPaRCS, that can be used by teachers to identify students that are in need of additional support and intervention and those that may qualify for access arrangements in their examinations. Feedback from teachers identified additional areas they would like to see assessed and indicated the need for digital administration that would allow scoring to be automated, saving time and reducing error. This presentation will discuss the development of a new and innovative digital psychometric test, SWIFT, designed to quickly identify school students at risk of learning difficulty or underachievement. The test has been designed for use in groups with students aged 11 to 18 years old and has norms from the UK and Ireland. The test consists of two alternate forms and assesses 5 core cognitive and academic areas: Spelling, Reading Fluency, Maths Fluency, Working Memory, and Processing Speed. The test was designed using a digital platform, EdPower, and preliminary data from over 1000 students has been analysed using SPSS and Rasch analysis studies conducted using Winsteps. The presentation will discuss the results of the data analysis along with some of the challenges associated with transitioning traditional pen and paper and individually administered tests into digital group assessments.

WEDNESDAY 3 JULY
Session 6.6
Topic: Translation of tests, psychological assessment instruments and survey questionnaire/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

722. Communication abilities in multilingual speakers with ADHD: insights from the Diagnostic interview for ADHD in adults.

Maria Garraffa
University of East Anglia and University of Oslo
Cecilie Rummelhoff, Franziska Köder
University of Oslo

Attention Deficit/Hyperactivity Disorder (ADHD) affects pragmatic language abilities, but so far pragmatic skills have only been investigated in the first language (L1). It remains unclear whether pragmatic communication in a second (L2) or third language (L3) is affected equally in multilingual speakers with ADHD. The study explored pragmatic abilities across different languages in 179 multilingual adults (91 adults with ADHD) with the Diagnostic Interview for ADHD in adults, a series of statements exploring with a self-rating 7 points Likert scale both inattentional and hyperactive/impulsive symptoms (DIVA, Kooij & Francken, 2010). For the inattention symptoms participants with ADHD reported more communicative difficulties than neurotypical participants across all three languages, and no modulation by language. For the statements concerning hyperactivity and impulsivity in communication, participants with ADHD also had generally lower ratings than neurotypical participants, with higher ratings in the L2 compared to the L1 and L3 compared to both L1and L2, meaning that participants with ADHD rated themselves as being less hyperactive and impulsive in conversations conducted in their second and third language. The results show that several communication abilities are affected significantly in adults with ADHD, with communication difficulties related to impulsivity (e.g., interrupting others or waiting one's turn) less pronounced in the L2 and L3 compared to the L1, potentially because with decreasing language proficiency, demands on speech planning and lexical retrieval increase. Clinicians working with individuals with ADHD should be aware that symptoms of ADHD might manifest differently in the languages spoken, encouraging assessment for ADHD in the language with the highest competence.

WEDNESDAY 3 JULY
Session 6.6
Topic: Translation of tests, psychological assessment instruments and survey questionnaire/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

734. Reliability and validity of the Test of Dyslexia for the Lithuanian speakers in second grade

Reda Gedutiene
Klaipeda University, Lithuania
Dovile Butkiene, Lauryna Rakickiene, Kestutis Dragunevicius, Grazina Gintiliene
Vilnius University, Lithuania

The dyslexia assessment tools have mostly been developed for English-speaking children. Due to this, the test for the native speakers of the Lithuanian language was developed. This test is based on the phonological deficit hypothesis of dyslexia and consists of thirteen subtests that measure rapid naming, word reading and spelling, pseudoword decoding and spelling, phonological awareness, and executive functions. This study aimed to examine the psychometric properties of the Test of Dyslexia for native Lithuanian speakers in the second grade. A representative sample comprises 192 second-graders (aged 8:4–9:4 years, 97 girls). The sample size for test-retest reliability was 33 children. In addition, 46 second-graders with reading and/or spelling disorders were assessed. Split-half reliabilities for subtests ranged from 0.81 to 0.99, for composite scales – from 0.91 to 0.99. Retest reliabilities for subtests ranged from 0.36 to 0.95, and for composite scores – from 0.61 to 0.83. The results of the confirmatory factor analysis support the construction of five composite scales (Reading Skills, Spelling Skills, Phonological Awareness, Rapid Naming, and Executive Functions) (factor loadings from 0.55 to 0.96). SEM analysis suggests that it is more appropriate to use the scores of the Reading and Spelling Skills scales to identify dyslexia, and to interpret them in the context of the other three scales and the scores of their subtests. Significantly lower composite scales scores of second-graders with diagnosed disorders (Cohen d from 1.25 to 1.87) support the concurrent validity of the test. The area under the curve for Reading Skills, 0.89 (95% CI 0.80-0.91) and for Spelling Skills, 0.90 (95% CI 0.86-0.94) show the good discriminative ability of these scales. The present study gives evidence of adequate psychometric properties of the Test of Dyslexia for native Lithuanian speakers in the second grade and shows the usefulness of the test for identification purposes.

WEDNESDAY 3 JULY
Session 6.7
Topic: Identifying biases by qualitative or quantitative methods

195. Reducing Evaluative Bias in Personality Assessment - Impact on Unboxing Neurodivergent Talent

Reda Gedutiene
Klaipeda University, Lithuania
Dovile Butkiene, Lauryna Rakickiene, Kestutis Dragunevicius, Grazina Gintiliene
Vilnius University, Lithuania

Objective - Traditional Big Five-based personality assessments are inherently biased through the use of imbalanced descriptors for low ends of each factor-polarity (e.g. Extraversion described as sociable, outgoing etc, and Introversion described as shy, withdrawn etc). This inherent bias leads to adverse impact whereby neurodivergent test-takers are subject to unfair criteria in high stakes contexts (e.g. recruitment, promotion opportunities), where ND individuals are more likely to score low on traditionally desirable Big Five factors (e.g. Extraversion, Openness) (Lodi-Smith et al, 2019). Sample - 293 adults globally, 49% self-identifying with at least 1 form of neurodivergence. Measures and Method - Lumina Spark personality assessment (Desson, 2017), AQ-10 ASD behavioural indicators (Allison et al, 2012), ASRSv1.1 ADHD Behavioural Indicators (Schweitzer et al, 2001), CAT-Q camouflaging scale (Hull et al, 2019). Behavioural scores of ASD, ADHD, and Camouflaging were correlated against both adaptive and maladaptive personality traits, identifying differential strengths between ND and neurotypical participants. Results - ASD individuals were found to score higher on measures of adaptive low-openness (Down to Earth) and low-Extraversion,(Introverted), displaying value over neurotypical participants in competencies involving Pursuing and Achieving Goals, Planning and Organising, Ensuring Accountability, and Gathering and Analysing Data. ADHD individuals were found to score higher on high-openness (Big Picture Thinking) and low-conscientiousness (Inspiration Driven), displaying value in competencies involving Adapting to Change, Agile Learning, Conceptualising Strategies, and Fostering Creativity. Implications - This approach of reducing evaluative bias in personality assessment provides a platform for reducing adverse impact against ND individuals in high-stakes workplace contexts, providing a shift towards a strength-based approach towards neurodivergence.

WEDNESDAY 3 JULY
Session 6.7
Topic: Identifying biases by qualitative or quantitative methods

294. The Inexact Science of Fairness Panel Reviews: Can They Make Assessments More Culturally Relevant?

Carina Fiedeldey-Van Dijk, Reuven Bar-On
Into Performance ULC

Response bias in testing is well documented in the literature. Test reports frequently feature two types: social desirability and inconsistency. Response bias refers to factors that lead test-takers to non-randomly diverge their responses from more accurate ones. This lowers the integrity of results, and risks making erroneous conclusions with costly implications. Response bias is expectantly addressed during test development, with careful item selection and judicious scoring. However, we also need to include such metrics in reports so that test users may directly evaluate the integrity of test results. We will introduce seven metrics that need to routinely appear in reports to facilitate accurate interpretation of results. We will show how test users can easily interpret and apply these metrics, with a discussion of their scientific rigor. The Bar-On Multifactor Measure of Performance (MMP) systematically examines the integrity of assessment results in its reports. We will explain its classification system of Valid, Credible, or Invalid, as determined by seven response-bias metrics. The cut-off points were statistically derived from the MMP's normative population (n=3831). Its reports offer practical guidelines when test takers might have scored outside any cut-off points. The MMP assesses two reliability and validity indices (self-image inconsistency and desirability [SIC, SID]) and three communality indices (completion time next to consideration time [CPT, CS]) and scale score range [RNG]). MMP reports also include two intercept indices (the perception of one's current level of performance [CLP] and attentiveness [ATT]). Its integrity metrics also prepare test users to select benchmarks for understanding and using their test results optimally. The stakes are high when leaders and decision makers use test results with uncertain accuracy for multiple applications at work and elsewhere. The MMP's approach is recommended for test authors and publishers to implement.

WEDNESDAY 3 JULY
Session 6.7
Topic: Identifying biases by qualitative or quantitative methods

476. The Inexact Science of Fairness Panel Reviews: Can They Make Assessments More Culturally Relevant?

Jessica Jonson
Buros Center for Testing - University of Nebraska-Lincoln
Sunhyoung Lee
Buros Center for Testing

Professional standards encourage the use of expert reviews to identify fairness issues in test development but evidence-based practices for ensuring these reviews effectively address cultural fairness issues are scarce. Professional standards and resources provide some guidance on conducting fairness reviews AERA, et al, 2014; ETS, 2016; ITC, 2018; Zieky, 2016. But, even racially diverse panels are ineffective in identifying insensitive or biased items Golubovich, et al., 2014; Reynolds, 2021; Sandoval & Mille, 1979. Additionally, Randall 2023 questions whether these reviews do more harm than good. Fairness reviews are widely used to collect fairness evidence in test development and potentially are a valuable source of evidence when used in a multiple and mixed-method approach to investigating test fairness Jonson, et al, 2019; Welch & Dunbar, 2022. However, more research on approaches and processes is needed to establish evidence-based practice. This presentation will discuss the results from a systematic analysis of fairness reviews documented in a sample of technical manuals for commercial tests. Our review found this documentation limited and lacking the details necessary to support the cultural fairness of the test. Details such as panelists' demographic background and culturally relevant expertise, instructions or training, and results were rarely provided. Future implications are a call for more research on the effectiveness of training and orienting review panelists, appropriate frameworks or criteria for making decisions, and how to structure the review process to improve the identification of culturally problematic items. Models for this type of research already exist for content alignment and standard-setting studies that also use expert panels. In addition, accountability for comprehensively documenting these reviews is necessary for test users and peers to evaluate the relevance processes used for addressing cultural fairness concerns.

WEDNESDAY 3 JULY
Session 6.7
Topic: Identifying biases by qualitative or quantitative methods

535. **Careless responding: trait or state?**

Inés Tomás Marco, Ana Hernández Baeza, Vicente González-Romá
University of Valencia / Spain
Anna Brown
University of Kent / United Kingdom
Clara Cuevas
University of Valencia / Spain

Theoretical/conceptual framework: Careless responding (CR) occurs when respondents fail to give sufficient attention to item content. Research has shown that CR behaviors are a source of bias, and it contributes to reducing the quality of the data (Podsakoff et al., 2012). It has been highlighted the need to prevent and manage CR (e.g., Arthur et al., 2021; Edwards, 2019; Ward & Meade, 2022). However, little is known about the nature of CR itself. Whereas some have approached it as a trait (Meade & Craig, 2012) others have argued that it is a state (Maniaci & Rogge, 2014), but no empirical evidence supports these claims. We respond to this gap by assessing the patterns of CR over time. Additionally, we test whether CR is a trait or state for the full population or there may exist subpopulations for which CR is a trait and others for which it is a state. Objectives: We analyze whether CR represents a stable response pattern over time (trait), or on the contrary, it is a transitory and fluctuating behavior (state). Additionally, we test whether there is heterogeneity of trajectories within the population and search for meaningful subpopulations on CR across time. Sample: 707 Spanish employees (50.4% men, aged between 21-59 years) who collaborated with a respondent panel. Methodology: We used a within-subject longitudinal design with 8 data collections spaced at 3-month intervals. The trajectory of CR over time was modeled by means of latent growth modeling (LGM), and latent class growth analysis (LCGA) using Mplus. Results: When analyzing the entire population, results suggest that CR represents a stable response pattern over time. However, four different subpopulations are identified: two stable groups (careful and careless individuals) and two for which the CR pattern changed over time. Implications: The study contributes to understanding the nature and dynamics of CR behavior. Study funded by the Spanish Ministry of Science and Innovation (PID2022-141339NB-I00).

WEDNESDAY 3 JULY
Session 6.7
Topic: Identifying biases by qualitative or quantitative methods

702. **Careless responding and DIF detection**

Ana Hernández
University of Valencia. Spain
María Dolores Hidalgo
University of Murcia. Spain
Inés Tomás
University of Valencia. Spain

The administration of tests, questionnaires and surveys in online mode through digital platforms and in digital environment is a common practice, which has led recently to the development of Guidelines for Technology-Based Assessment (ITC, 2022). Among the problems arising from this type of assessment is the probability of generating careless responses, and the effect that the lack of attention of examinees when answering questionnaires produces on the estimation of the psychometric properties of the tests. Taking it a step further, the question arises: How does a careless response pattern affect the correct identification of items with Differential Item Functioning (DIF)? Thus, the objective of the present research is to analyze the behavior of different DIF methods in the presence of careless responding using a simulation study. Several factors are manipulated such as number of DIF and percentage of participants exhibiting careless responses. Other factors such as number of test items (5, 8 or 10 items), type of DIF (uniform), and sample size are considered. These factors are set to conditions more commonly encountered in practice. Additionally, an empirical example derived from a survey questionnaire regarding work quality, well-being and health is analyzed. The effects of careless response in DIF detection in an applied context are shown. Implications for practice are discussed.

WEDNESDAY 3 JULY
Session 6.8
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

123. Integrating Validity Evidence for a Comprehensive Protocol to Assess Competencies in Incoming University Students

Juan F. Luesia, Milagrosa Sánchez-Martín
Universidad Loyola Andalucía
Isabel Benítez
University of Granada

Current educational paradigms, such as the 21st-century skills or the "quality education" Sustainable Developmental Goal, emphasize the importance of extending the nature of the competencies assessed and of involving diverse university stakeholders in evaluations. New evaluations should also provide validity evidence to support their intended use and scores' interpretation. The present study aims to integrate various sources of validity evidence obtained to support the intended use of a comprehensive protocol (CompassIn) created to assess academic competencies in incoming university students, through 25 measures (four cognitive tasks, and 21 non-cognitive measures). Three sources of validity evidence were gathered based on: test content (through a mixed design combining outputs from a systematic review, focus groups with experts, and a large-scale assessment); internal structure (through reliability and network analysis); and relationship with other variables (through correlation, regression and sensitivity analyses). Different stakeholders participated in different phases of the study, from teachers or university staff to a sample of 607 university students. Results indicated enough evidence to support the intended use of the CompassIn. The first phase provided empirical evidence to support a theoretical model of academic competencies associated with university success. Secondly, the administration of the protocol confirmed the adequacy of the measures in terms of psychometric properties (alpha and omega ranged between .74 and .94, confirming the reliability) and dimensionality, as the network analysis confirmed the overlap between the theoretical model and the variables measured. Finally, relationships between CompassIn scores and academic success were as expected. Integrating results from different phases highlighted the benefits of developing mixed designs for validation studies. Advantages and limitations will be discussed.

WEDNESDAY 3 JULY
Session 6.8
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

301. MARKO-D a cross-cultural tool for early mathematics assessment

Victoria Espinoza, Ricardo Rosas Pontificia
Universidad Católica de Chile

Mathematical competence plays a fundamental role in students' adequate academic and professional development, as it allows them to understand and process numerical information appropriately. The assessment of early mathematical skills is crucial, so it is essential to have reliable assessment tools that can be applied in a variety of assessment contexts. MARKO-D is a test that assesses early mathematical skills, it was developed in Germany, but has been adapted to different languages and cultural contexts. The MARKO-D Chile test was constructed on the basis of the original German test, which allowed empirical validation of the mathematical concept development model underlying the test. The test was adapted and validated in Chile with a sample of 293 students from pre-kindergarten to second grade. A Rasch analysis was conducted to test the empirical functioning of the theoretical model underlying the MARKO-D test. Content validity is based on the soundness of the theoretical model on which the test is developed and also on the fact that all test items were designed, evaluated and subsequently adapted by experts. To establish criterion validity, the results of a sample of the total group were correlated with the results of the WISC-V test, specifically with the subtest of arithmetic (convergent) and construction with cubes (divergent). The correlation with the convergent construct (P=.708, p

WEDNESDAY 3 JULY
Session 6.8
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

627. Does PISA Literacy Assessment Provide Fair Comparisons Across Participating Countries?

Semih Topuz
Baskent University
Semirhan Gokce
Nigde Omer Halisdemir University
Giray Berberoglu
Baskent University

The translations may affect the meaning and role of words, sentences, and passages, the content of items, and the abilities assessed by the items (Ercikan, 1998). Cross-cultural fair comparisons require equivalency in scale characteristics across languages and countries. The PISA follows a very detailed and strict preparation and analyses to ensure the culture and language free comparisons among participating countries. However, there could be inequalities in the scale characteristics when language and country interact with each other even in tests which relay on limited use of language such as mathematics. Comparability of test scores in TIMSS has also been found to be problematic across language and cultural differences(Gökçe et al., 2021). Comparisons across the countries are common practice in the international assessment programs. Thus, the fairness of these comparisons becomes a critical issue when there are language and cultural differences among the participating countries. Grisay et al. (2009) investigated the equivalence of the reading assessments used in PIRLS 2001 and PISA 2000 reading. They found higher DIF magnitudes for non-Indo-European language translations. Since PISA tests are basically language based assessment, having large difference for the non-Indo-European languages is expected. Thus, in the present study, how linguistic and cultural diversity may affect the performance of the students on literacy measures used in PISA will be investigated. The three combinations of languages and countries such as, same language but different countries, same countries but different languages, and different languages and different countries are considered to evaluate consistency of students' responses by using DIF analysis. The magnitude of DIF as well as the difference between test characteristic curves are compared across three combinations of languages and countries. The results will be interpreted in line with country and language interactions.

WEDNESDAY 3 JULY
Session 6.8
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

646. More Than Grades: Investigating an Assessment Tool Measuring Socio-emotional Competence and Their Links to Academic Outcomes

Aira Vener Peregrin, Katerina Angelica Javier
Global Resources for Assessment Curriculum and Evaluation, Inc

Socioemotional competencies (SEC) are increasingly recognized as important components of academic success beyond cognitive ability and achievement. The value of establishing SEC empirically as part of a holistic, successful education is therefore imperative. The framework measures five interconnected components of socioemotional competency (self-awareness, self-management, social skills, academic self-regulation, and grit) that interact to predict academic success. This study investigates the use of a measure of SEC as a predictor of academic performance in Filipino students at the primary (n = 200), intermediate (n = 200), and Junior High School (n = 250) levels. Multiple linear regression was performed on the student's scores on a standardized achievement test and the SEC measurement tool. Results showed a significant but mild correlation between the achievement test scores and the combined scores from the SEC measurement tool at the primary (r = .316, .341, .256 for English, Math and Science respectively; p = < .001), intermediate (r = .198, .271, .163; p = .005, < .001, .021) and junior high school levels (r = .176, .182, .255; p = .005, .004,

WEDNESDAY 3 JULY
Session 6.8
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

776. Questionnaire-taking motivation and score validity: Satisficing in the PISA 2018 student questionnaire assessed through response process data

Hanna Eklöf, Erik Lundgren
Umea University, Umeå University

Lack of motivation to respond carefully and truthfully to questionnaire items, in survey research known as satisficing, can impact the validity of conclusions drawn from survey responses. Although the issue of careless responding to questionnaires in large-scale educational assessment such as TIMSS and PISA has received some attention, rather little is known about the occurrence of satisficing, or about the impact of satisficing on item scores and interpretation of item scores from these questionnaires. Hence, the 'questionnaire-taking motivation' of students seems like a relevant task to explore. The present study provides an example of such an exploration, where response process data from the computer-based PISA 2018 student questionnaire were used to estimate the prevalence of satisficing by modeling a) response times and b) response times and dependencies in subitem responses jointly. Overall results indicated that most students were motivated to provide valid responses to most items in the PISA student questionnaire. Yet, questionnaire items in PISA are of varying complexity and some of the more complex items did not fit well to the model applied. Further analyses of these items suggested a high prevalence of satisficing (aberrant responding) and that this caused bias in the scores derived from the items. It is concluded that response process data can be valuable indicators for assessing survey response quality, and also that the motivation of students to respond to questionnaire items as well as the cognitive load of such items should be considered when developing and analyzing large-scale questionnaires.

WEDNESDAY 3 JULY
Session 6.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

112. Construct Validity of International WISC Versions: Informing Evidence Based Assessment

Gary Canivez
Eastern Illinois University/USA

Wechsler intelligence scales are among the most frequently used individual measures of cognitive abilities world-wide (Evers et al., 2012; Georgas, van de Vijver, Weiss, & Saklofske, 2003; Lichtenberger & Kaufman, 2009), with numerous translations, adaptations, and standardizations. Publication of the WISC–IV (Wechsler, 2003) and WISC–V (Wechsler, 2014) led to numerous international versions that require independent psychometric evidence to support interpretation of provided scores and comparisons. It is important for test users to select and interpret instruments for which there is adequate supporting psychometric evidence. This symposium presents a collection of four papers detailing independent psychometric evaluations of the WISC–IV and WISC–V in four cross-cultural contexts (Brazil [WISC–IV], Canada, Korea, Australia/New Zealand [WISC–V]). Paper 1 reports independent structural and reliability analyses of the Brazilian WISC–IV standardization sample using hierarchical exploratory factor analyses (HEFA) and confirmatory factor analyses (CFA). Paper 2 presents independent structural validity and reliability analyses of the Canadian WISC–V with an indigenous sample using HEFA and CFA with comparisons to the standardization sample. Paper 3 presents the structural validity and reliability analyses of the Korean WISC–V using HEFA and CFA with the standardization sample. Paper 4 reports on independent structural validity and reliability analyses of the Australia/New Zealand WISC–V with the standardization sample. In addition to HEFA and CFA, all report results from multiple methods for determining the number of factors to extract, exploratory graph analysis, and model-based reliability/validity estimates to facilitate consideration of viable score interpretation; and replication of independent research substantially challenging publisher claims. These are essential for test users to "(a) know what their tests can do and (b) act accordingly" (Weiner, 1989, p. 829).

WEDNESDAY 3 JULY
Session 6.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

165. Construct Validity of the Canadian WISC–V with an Indigenous Sample: Hierarchical EFA and CFA (Validity and fairness in cross-cultural testing, psychological assessment and survey research)

Gary Canivez
Eastern Illinois University
Ryan McGill
William & Mary/USA
Jessica Hanson
Champaign School District Unit 4/USA
Merril Dean
Private Practice/Canada

The Canadian WISC–V (WISC–V CDN; Wechsler, 2014a) provides a five-factor structure promoted by the publisher but Watkins et al. (2018) noted several WISC–V CDN psychometric problems of the proffered model (poor discriminant validity and local fit problems). Results showed best fit for bifactor structure with four group factors (Wechsler Model) and model-based reliability/validity estimates showed most of the variance in the group factors was due to general intelligence. Strong support was only observed for the measurement of g. A lack of research investigating WISC–V CDN construct validity with indigenous children is addressed by this study that assessed the factor structure using EFA and CFA with a sample (N=296) of indigenous children in Canada. EFA illustrated inadequacy of all but the 4-Factor model with no separate Visual Spatial and Fluid Reasoning factors (no salient subtest loadings on Factor 5). Higher-order EFA with SL transformation and model-based reliability/validity estimates found strong measurement of g but poor unique measurement of group factors (VC, VS [PR], WM, PS). CFA indicated models with four or five group factors all fit well based on model fit statistics, but both higher-order models contained standardized paths of 1.0 from g to a group factor, indicating inadequacy. Global fit statistics indicated best model fit was the bifactor with four group factors, but local fit problems required dropping nonsignificant group factor paths. The bifactor model with five group factors was plausible but also contained local fit problems. After dropping the nonsignificant group factor paths for Fluid Reasoning this was the most plausible final model matching WISC–V CDN EFA results with Canadian indigenous youth and Watkins et al. (2018) WISC–V CDN CFA of the standardization sample. Like EFA, CFA model-based reliability/validity estimates found strong measurement of g, but poor unique measurement of group factors (VC, VS [PR], WM, PS).

WEDNESDAY 3 JULY
Session 6.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

168. Construct Validity of the Korean WISC-V: Hierarchical EFA and CFA (Validity and fairness in cross-cultural testing, psychological assessment and survey research)

Gary Canivez
Eastern Illinois University/USA
Ryan McGill
William & Mary/USA
JungSu Oh
Eastern Illinois University/USA
Juyeon Lee
Korean Educational Development Institute/Republic of Korea

The Korean WISC-V (WISC-VK; Kwak & Jang, 2019) includes five factor scores and a higher-order Full Scale IQ based on the publisher preferred five-factor structure; however, numerous independent WISC-V examinations with U.S. and international versions repeatedly show support for only four first-order factors (Canivez et al., 2016, 2017, 2018, 2021; Lecerf & Canivez, 2018, 2021; Fenollar-Cortes & Watkins, 2018; Watkins et al., 2017) with Visual Spatial and Fluid Reasoning merged. Presently, there is no technical manual reporting psychometric properties of the WISC-VK and no independent assessment of the latent factor structure, so the present study assessed the WISC-VK standardization sample (N=2,257) subtest correlation matrix using EFA and CFA and model-based estimates for reliability and dimensionality. No factor extraction criterion suggested five factors except publisher claim. Extraction began with five factors and iteratively reduced by one to explore alternate solutions. The five-factor extraction showed the fifth factor contained no salient subtest pattern coefficients so was inadequate. The four-factor extraction included salient subtest loadings on theoretically consistent factors, except Picture Concepts (no salient loadings). Second-order EFA and Schmid and Leiman transformation yielded 64.2% of explained common variance (ECA) by general intelligence, and group factors provided only 5.5%–11.3% additional ECA. CFA fit statistics for all hypothesized models illustrated only models with four or five group factors were adequate or well-fitting but all had local fit problems. Bifactor models were best fitting and meaningfully better than higher-order models, but local fit problems in bifactor models required dropping non-significant path coefficients and model re-estimation. Local fit problems were particularly related to inadequacy of Fluid Reasoning and model-based estimates for reliability and dimensionality supported only the g factor as also found in EFA.

WEDNESDAY 3 JULY
Session 6.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

197. Construct Validity of the Australia/New Zealand WISC–V: Hierarchical EFA and CFA (Validity and fairness in cross-cultural testing, psychological assessment and survey research)

Ryan McGill
William & Mary
Gary Canivez
Eastern Illinois University
Soheil Afshar
Macquarie University
Marley Watkins
Baylor University

The Australia/New Zealand WISC–V (WISC–VANZ; Wechsler, 2016) includes five factor scores and a higher-order Full Scale IQ based on the publisher preferred five-factor structure; however, numerous independent WISC–V examinations with U.S. and international versions repeatedly show support for only four first-order factors (Canivez et al., 2016, 2017, 2018, 2021; Lecerf & Canivez, 2018, 2021; Fenollar-Cortes & Watkins, 2018; Watkins et al., 2017) with Visual Spatial and Fluid Reasoning merged. Presently there are no independent structural validity investigations for the WISC–VANZ so the present study examined the WISC–VANZ factor structure with the standardization sample (N=528) subtest correlation matrix using EFA and CFA and model-based estimates for reliability and dimensionality. No factor extraction criterion suggested five factors except publisher claim, so extraction began with five factors and iteratively reduced by one to explore alternate solutions. Five-factor extraction showed the fifth factor contained only two salient subtest pattern coefficients, one cross-loaded and one was theoretically inconsistent. Four-factor extraction included all subtests with salient loadings on theoretically consistent factors, except Picture Concepts. High factor correlations required second-order EFA and Schmid and Leiman transformation and the g factor had 66.9% explained common variance (ECA) and group factors provided 7.1%–9.7% additional ECA. CFA fit statistics illustrated only models with four or five group factors were adequate or well-fitting, but all had local fit problems. Models with four group factors were superior to five group factors (improper solutions due to negative residual variance estimates [FR-g loading >1.0]). Bifactor and higher-order representations with four group factors were not meaningfully different. Model-based reliability estimates showed adequate portions of true score variance for g, but inadequate unique variance among the group factors.

WEDNESDAY 3 JULY
Session 6.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

331. Construct Validity of the Brazilian WISC-IV: Hierarchical EFA and CFA (Translation of tests, psychological assessment instruments and survey questionnaire)

Solange Muglia Wechsler
Pontificial Catholic University of Campinas-Brazil
Gary Canivez, Ryan McGill, Nicholas Benson
United States

The Brazilian WISC-IV is the only validated measure in the country so far. Its adaptation to Brazil was based on the same four first-order factors and higher-order g structure factor as the U. S. and other international versions. Although there was a proposition from other authors on 5 factors for the WISC-IV, there was no independent assessment of its structure with the normative sample. This study investigated the Brazilian WISC-IV factor structure with the standardization sample (N=1,863) subtest correlation matrix using EFA (with oblique rotation) and CFA and estimates for reliability and dimensionality. Extraction criteria indicated 1-5 possible factors. The five-factor extraction was inadequate (no salient fifth pattern coefficient) but the four-factor extraction yielded salient loadings for all subtests aligned with the WISC-IV based dimensions. Factor correlations (.427-.785) necessitated second-order EFA and Schmid and Leiman transformation, where variance estimates revealed a g factor accounted for 68.3% of the explained variance (ECV). In contrast, the four group factors provided 2.4%-12.3% ECV. The g factor omega-hierarchical coefficient was large (.824) but the group factors omega-hierarchical subscale coefficients were small (.76-.498) and below the minimum criterion. CFA indicated the Wechsler based bifactor model fits best, but not meaningfully better than the CHC based bifactor model. Both bifactor models were meaningfully better than the Wechsler and CHC higher-order models. The g factor was essentially unidimensional, and the four (Wechsler) or five (CHC) group factors omega coefficients in all models were low and indicated poor unique variance, and thus questionable interpretative value like EFA.

WEDNESDAY 3 JULY
Session 7.1 SYMPOSIUM
Topic: Innovations in test development

276. Tech and Talent Synergy: Innovation for DEI, Individual, and Organizational Performance

Richard Justenhoven, Maximilian Jansen
Welliba

The world of work has been characterized by frequent change resulting in acronyms like VUCA (Millar et al., 2018). These changes are often closely related to technological advancements, requiring research and practice in IO psychology to continually evaluate and adapt current practices. This symposium explores different aspects of the complex interplay of technology and applicant-/employee-employer dynamics. Particular attention is given to individual and organizational outcomes such as assessment and job performance, employee experience (EX) as construct on the employee-employer relationship (Morgan, 2017), as well as current challenges and opportunities in diversity, equity, and inclusion (DEI) across the employee lifecycle. All presentations in this symposium feature quantitative data analyses with samples predominantly recruited from the working population, interpreted against a backdrop of theoretical foundation and practical implications, to provide comprehensive insights in the presenters' respective areas of expertise. Focusing on psychometric assessments, Asmaro and Englund present a novel cognitive assessment using latest design principles and technologies to reduce subgroup differences without compromising construct validity. This is complemented by Leutner sharing insights on the psychometric properties of image-based tests and their perception by neurodiverse applicants. Looking at employees, Kurz presents a co-validation study of a new model of job success factors featuring data on a job success scale and a variety of personality and competency constructs. Building onto this, Preuss et al. share a novel approach to ongoing measurement of EX over time and data on the links between EX and organizational outcomes. By highlighting these technologies, methodologies, constructs, and outcomes, this symposium aims to combine a snapshot of current practitioner research, with practical recommendations for anyone looking to hire or retain and develop talent.

WEDNESDAY 3 JULY
Session 7.1 SYMPOSIUM
Topic: Innovations in test development

304. Perceptions of Image- and Questionnaire- Based Personality Measures Among Neurodivergent Adults (Translation of tests, psychological assessment instruments and survey questionnaire)

Franziska Leutner, Airlie Hilliard
Goldsmiths

Algorithmic recruitment tools are increasingly being used to enhance candidate experience, identify qualified applicants, and rapidly measure psychological constructs such as cognitive ability and personality. Much of the research so far has concerned the validity of the tool and user experience, with some also examining fairness perceptions and user experience associated with novel test formats such as game-based assessments and video interviews. However, much of the existing research focuses on the performance and perceptions of algorithmic recruitment tools among general, typical populations, with a lack of research on how they are perceived among and perform for neurodivergent populations. As such, an initial qualitative study was conducted that interviewed neurodivergent adults, particularly focusing on those with ADHD, dyslexia, and autism, on their perceptions of traditional and algorithmic pre-employment tests. Findings indicated that perceived and actual barriers associated with pre-employment tests were not unique to either test format but were a significant source of stress and anxiety, which could affect performance. However, enhancements such as gamification and considerations of user interface and graphics were suggested as ways to reduce some of these barriers and were more associated with algorithmic tools than traditional formats. This study, therefore, builds on the findings of the qualitative study to quantitatively measure reactions to an algorithmically scored image-based assessment of personality and creativity designed for use in selection compared to questionnaire-based equivalent among neurodivergent respondents recruited via Prolific Academic. Within and between group analysis will be conducted to investigate whether either format is perceived more positively than the other among adults with a diagnosis of ADHD, dyslexia, or autism, as well as whether the image-based format is preferred among particular neurotype. Perceptions will also

WEDNESDAY 3 JULY
Session 7.1 SYMPOSIUM
Topic: Innovations in test development

279. Advancing Fairness by Reducing Subgroup Differences with a New Logical Assessment (Innovations in test development)

Mats Englund
Fairsight/Sweden
Fredrik Asmaro
Fairsight/Norway

In the evolving landscape of industrial and organizational psychology, fairness and Diversity, Equity, and Inclusion (DEI) in talent selection are paramount. Addressing these topics include improving candidates' subjective experience during selection processes as well as the objective outcomes of those processes (e.g., less adverse impact). Using engaging and easy-to-understand assessments helps to create a positive assessment experience for candidates. To the extent that psychometric methods result in lesser subgroup differences, practitioners are better able to handle the validity-diversity dilemma in selection processes (Pyburn, Ployhart, and Kravitz, 2008), resulting in less compromise between short-term and long-term goals of organizations and society at large. This presentation introduces a novel logical assessment tool for measuring general logical reasoning (i.e., high "g-loading"), which was specifically designed to address the critical need for reduced subgroup differences in test scores while maintaining candidate engagement and construct validity. Utilizing modern technology and UX-design, improvements of typical assessment design were made, resulting in an interactive approach to classic matrix-type reasoning assessments. Initial studies were conducted with samples from the working population recruited via an online crowdsourcing platform. The results so far have been promising; main findings from the studies include positive candidate reactions and subgroup differences smaller than what is typically found for cognitive ability (e.g., Roth et al. 2001), while maintaining construct validity.

WEDNESDAY 3 JULY
Session 7.1 SYMPOSIUM
Topic: Innovations in test development

296. Continuous Insights, Lasting Impact: The interaction of mindset and context factors in shaping Employee Experience (Innovations in test development)

Maximilian Jansen, Achim Preuss, Richard Justenhoven
Welliba/Germany

Recent years have seen substantial shifts not only in technology, but also in employee-employer dynamics and labour market trends. These have been marked by an increasing pervasiveness of concept of Employee Experience (EX; Justenhoven et al., 2023) in HR contexts. Though the use of the term employee experience in practice is not entirely consistent, it can be broadly defined as "the relationship between an employee and the organization" (Morgan, 2017, p. 7). From this perspective, EX is relevant throughout the employee lifecycle and can thus impact a variety of individual and organisational outcomes (Preuss & Justenhoven, 2023). Another consequence of this view is that EX is likely to be influenced by a range of factors. This is reflected in the EX model and measurement approach this presentation introduces. Drawing on the framework of Self-Determination-Theory (SDT) our model emphasises the interaction of an individual's mindset and contextual factors in their work environment, as well their fluctuations over time. This is reflected in the CadaMint measurement approach based on and named after continuous adaptive micro interactions, designed to encourage frequent data collection in short sessions to allow accurate modelling of individual and aggregated trends over time. CadaMint was validated against eight business outcomes including intention to quit, abseentism risk, and trust in one's employer, demonstrating the relevance and value of managing and improving EX for organisations. A continuous measurement approach allows for significantly more detailed views of trends and changes over time on team, department, or organization levels. The ability to identify trends and dynamics can be crucial in effectively managing or even anticipating challenges related to EX. The density of data points over time also allows for high degrees of personalisation of measurements and outputs based on user profiles, thus also generating value for individuals.

WEDNESDAY 3 JULY
Session 7.1 SYMPOSIUM
Topic: Innovations in test development

287. Job Satisfaction and Performance: What are the Great 8 Drivers of Job Success? (Innovations in test development)

Michele Guarini
Denmark

The World of Work (WoW) model (Kurz & Bartram, 2002) postulated that performance and affect are impacts of person and environment interaction with the 'Great 8 Competencies' as constructs underpinning job success. The objective of this paper is to explore a two-item 'Job Success' scale and its relationship with a newly developed Great 8 Success Factors model. In a co-validation study covering numerous personality and competency constructs, 466 professionals and managers completed 758 items that were rated on a 9-point scale adopted from the TDA (Lewis, 1992). The items 'I am overall effective in my current job' and 'I am overall satisfied with my current job' were designated as a global measure of Job Success. Aligned Personality Factors (PF48) predictor and Competency Factors (CF48) criterion measures with 48 facets grouped into 8 factors were developed. The resulting Success Factors were correlated with the Job Success scale and the individual items. Positive correlations between all Great 8 measures and the Job Success scale and its constituent items were expected and obtained with factor correlations ranging from .14 to .39. The average Competency Factors (CF48) correlation with Job Success was .23, with the effectiveness item .28 and with the satisfaction item .18. The average Personality Factors (PF48) correlation with Job Success was .27, with the effectiveness item .28 and with the satisfaction item .22. CF48 mean reached .41 and PF48 Total .40 for Job Success. For the effectiveness item 84% of the ratings were positive, 4% 'Unsure' and 12% negative whereas for the satisfaction items the respective values were 70%, 4% 'Unsure' and 26%. The scale internal consistency reliability Alpha was .77 with a correlation of .64 between the items. The implication of the research is that the Job Success scale is concise but viable with job satisfaction and effectiveness being closely related and consistently linked to the Great 8 Success Factors.

WEDNESDAY 3 JULY
Session 7.2 SYMPOSIUM
Topic: International assessment

555. Challenges and Strategies: Measurement and Testing Associations in Focus

Nekane Balluerka (University of the Basque Country), Nigel Evans (European Federation of Psychological Associations), Ana Hernández (European Association of Methodology), Albert Sesé (Spanish Association of Methodology), Steve Sirecci (International Test Commission)

WEDNESDAY 3 JULY
Session 7.3 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

104. Bridging International Perspectives on Socioemotional Learning and Assessment

Javier Suárez-Álvarez
University of Massachusetts Amherst

Nations globally continue to grapple with persistent disparities in educational outcomes, which have been exacerbated by physical and mental health difficulties stemming from the COVID-19 pandemic.

Cognition and emotion are interrelated components of learning, and their interconnected development starts during early infancy and continues throughout the lifespan. While social and emotional skills can be taught, their assessment remains challenging due in part to varying frameworks and research methods. While different frameworks and research methods can help accommodate cultural and linguistic barriers, they can also lead to highly heterogeneous results when evaluating the effectiveness of Social and Emotional Learning (SEL) interventions. In this symposium, we will share research experiences from diverse countries about designing and measuring socio-emotional learning in varied educational settings. The goal is to examine the theoretical, methodological, and practical challenges of assessing socioemotional skills and discuss how to improve comparability and fairness across cultural and linguistic contexts. The first presentation will describe AVANZO and Conociendo mis Logros, a comprehensive socio-emotional assessment of El Salvador's education system. The second presentation will illustrate top-of-the-art social-emotional assessment examples from Brazil. The third presentation will illustrate the translation and validation of the Behavioral, Emotional, and Social Skills Inventory (BESSI) in Spain. The fourth presentation will illustrate how the OECD's Study on Social and Emotional Skills assesses these skills internationally to inform evidence-based policymaking. Beatrice Rammstedt, a leading researcher in survey methodology, psychological assessment, and large-scale assessment, will discuss the papers. Sufficient time will be allotted for audience Q&A.

WEDNESDAY 3 JULY
Session 7.3 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

133. Assessing Socio-emotional Skills in Spain: Development and Validation of the Spanish Version of BESSI (Translation of tests, psychological assessment instruments and survey questionnaire)

Álvaro Postigo
University of Oviedo
Covadonga González-Nuevo, Jaime García-Fernández
University of Burgos
Rubén Fernández-Alonso, Marcelino Cuesta
University of Oviedo

Social, emotional and behavioral skills comprise a broad set of abilities essential for establishing and maintaining relationships, regulating emotions, selecting and pursuing goals or exploring new stimuli. These have shown an important impact on people's lives. To improve their assessment in Spain, the aim of the present study was to adapt and validate the Behavioral, Emotional and Social Skills Inventory (BESSI) in the Spanish adult population. The BESSI measures 32 facets of 5 domains through 192 items. The psychometric properties of the Spanish version of the BESSI were studied (internal consistency, test-retest reliability, evidence of validity in terms of internal structure and in relation to other variables). The results show that, in general terms, each of the facets is unidimensional, with adequate internal consistency and stability of its scores and with adequate evidence of convergent validity in relation to the Big Five. In turn, the structure at the domain level is very similar to that of the original English version. The Spanish version of the BESSI has shown adequate psychometric properties, so it can be used in the general Spanish population to assess social, emotional and behavioral competencies in adult population.

WEDNESDAY 3 JULY
Session 7.3 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

200. Validation and scaling of the OECD Survey on Social and Emotional Skills (SSES) (International assessment)

Elena Govorova, Elena De la Guía
2E Estudios & Evaluaciones

Conceptual framework: The Survey on Social and Emotional Skills (SSES) is a cross-national study promoted by the OECD and conducted in various countries and sub-national entities (sites). The survey focuses on evaluating the social and emotional skills of 15-year-old students (optionally 10-year-olds). The student assessment encompasses 16 constructs, resulting in 15 scales derived from the Big-Five model, along with an additional skill scale for achievement motivation. Objectives: The primary goals of the data validation and scaling of SSES data were twofold: to assess the psychometric properties of SSES scales and to compute scales for each of the 16 constructs outlined in the framework. Sample: The Main Survey was conducted during Spring and Autumn 2023 across 16 sites. The total number of 70.114 students participated in the second round of SSES Round 2 representing over 2,5 million of students. Methodology: Firstly, psychometric properties of items and scales were assessed through confirmatory factor analysis, evaluating constructs and conducting multiple-group confirmatory factor analysis to ensure measurement equivalence across various groups (age cohorts, participating sites, gender groups, and rounds). The second step involved data scaling using the IRT Generalised Partial Credit Model, with subsequent score generation for survey participants. Results: The results indicate that all 16 SSES scales demonstrated acceptable psychometric properties concerning internal consistency reliability and unidimensionality for the pooled international dataset. While metric invariance for gender and cohort groups was met for all scales, scalar invariance was achieved by only a subset of the 16 scales. Implications: The implications of the robust and high-quality international dataset of SSES are substantial. It provides a solid foundation for shaping effective educational policies and interventions designed to enhance the social and emotional skills of students globally.

WEDNESDAY 3 JULY
Session 7.3 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

161. Assessing Social-emotional Skills using SENNA: Development, Psychometrics and Validity (Validity theory in testing, psychological assessment and survey research)

Ricardo Primi
Universidade São Francisco & EduLab21 Ayrton Senna Institute
Oliver P. John
University of California, Berkeley, USA
Filip De Fruyt
Ghent University, Ghent, Belgium
Daniel Santos
University of São Paulo, Ribeiro Preto, Brazil

In the contemporary educational landscape, fostering social-emotional skills (SEMS) is pivotal for preparing students to navigate a complex and unpredictable world. SEMS, shaped by biological and environmental influences, are manifested in consistent thought, feeling, and behavioral patterns. They evolve through various learning experiences, impacting life outcomes. Recent initiatives have sought to consolidate diverse SEMS frameworks into comprehensive models. Our study presents SENNA, an assessment tool developed by the Institute Ayrton Senna, designed to evaluate a broad SEMS model. This model aligns closely with the Big Five personality domains, encompassing Self-management, Engaging with Others, Amity, Negative Emotion Regulation, and Open-mindedness. SENNA employs a self-report inventory with 162 items across these domains, organized into 18 facets. Study 1, involving over 30,000 Brazilian students aged 11-18, employed exploratory structural equation modeling. This confirmed the expected domain loadings with high congruency coefficients and satisfactory internal consistencies. Study 2, with 1041 students, compared SENNA against other SEMS measures like SAL, MESH, and Character Growth Card, demonstrating robust construct and concurrent validity. The final study examined SENNA's criterion validity in predicting standardized test scores in mathematics and language. Utilizing a sample oof 12,987 students from 425 São Paulo State schools, the results highlighted SEMS facets' significant contributions to academic performance, enhancing variance explanation in test scores. This presentation highlights SENNA's effectiveness as a comprehensive, reliable, and valid tool for assessing SEMS in educational settings, contributing significantly to understanding and measuring socio-emotional learning in Brazil.

WEDNESDAY 3 JULY
Session 7.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

91. Integrating Transformer Models for Culturally Adaptive Trait Classification

Tales Marra, Emeric Kubiak
AssessFirst

Objective. Content validity is essential in personality assessment, particularly in the context of adapting evaluation instruments to minority languages. Progress in NLP allow to use transformer models for the task of trait classification from texts, outperforming humans on the trait prediction task and fostering more efficient early-stage content validity evaluation. This progress opens doors for more accurate assessment tools in minority languages. Our research builds upon Fyffe et al. (2023) by training our model on a more extensive pool of items, and enhancing our understanding of the model's predictions. Method. First, we compiled a training dataset combining the original Fyffe et al. (2023) dataset with added items from reputable open-source assessments (e.g., HEXACO-100), culminating in 3,905 items measuring 6 traits. Second, items were encoded as digital vector. Third, our model was trained at identifying the relationships between items characteristics and traits. We used DeBERTa (He et al. 2021). We applied our model to a neutral and validated personality assessment (Kubiak et al., 2022) with 75 forced-choice items. An attention analysis helped to understand which parts of the items contribute most to the final prediction. Results. The model demonstrated remarkable accuracy (.92), precision (.91), recall (.92), and F1-score (.91) across traits. Results were slightly lower for Extraversion (F1-score = .81) and Humility (F1-score = .82). Confirmed by the embedding analysis, some items from Humility were classified in Extraversion. Attention analysis show the relative importance of some markers in the prediction, like 'others' for Agreeableness, 'ideas' for Openness, or 'work' for Conscientiousness. Conclusion. By analyzing how specific linguistic markers correlate with personality traits in different culture, this methodology could offer a nuanced approach to understanding cultural variations in language use.

WEDNESDAY 3 JULY
Session 7.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

249. Beyond Surveys: Multi-Modal Measures of 21st Century Skills Across Classrooms

Therese Hopfenbeck
Assessment and Evaluation Research Centre, University of Melbourne, Australia
Samamtha-Kaye Johnston
Department of Computer Science, University of Oxford, UK
Juliet Scott Barrett
Department of Education, University of Oxford
Tracey Denton - Calabrese
Oxford, UK
Joshua McGrane
Assessment and Evaluation Research Centre, University of Melbourne, Australia

The validity of international survey data on 21st century skills like creativity, curiosity, and critical thinking has been questioned. Given the biases in self-reports, this study leveraged multi-modal measures across nine countries' classrooms. Video recordings, interviews, and tests assessed creativity and curiosity in Denmark, Germany, Ghana, France, India, Italy, Netherlands, Norway, and Sweden. This study explored how 22 primary school teachers in these countries facilitated self-regulation to encourage creativity across 19 classrooms. In 2021, 46 videos were remotely gathered during the pandemic. Teachers and students were interviewed on their experiences with in-class self-regulation strategies and links to developing creativity. Through reflexive analysis, researchers categorized practices tied to deploying self-regulation to further creativity. Findings showed setting goals, planning, time management and accounting for individual differences. The results showcase this methodology's value for responsive pedagogy and diverse, achievable international classroom research. Furthermore, these videos could inform efforts to validate survey instruments relying on self-reported constructs such as different approaches to learning, curiosity, creativity and critical thinking.

WEDNESDAY 3 JULY
Session 7.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

252. Evaluating the Quality of University Curriculum: A Theoretical Framework, Methodology, and Empirical Analysis

Wen Wen, Lu Zhou, Yizhuo Chen
Tsinghua University

I. Introduction The teaching and learning of advanced knowledge distinguish university curriculum from basic education. This study focuses on the essential nature of university curriculum as the primary institutionalized carriers of advanced knowledge in higher education. II. Theoretical Framework Referring to international and Chinese studies, this research identifies four key characteristics of university curriculum: knowledge-specification, cognitive-challenging, research-oriented, and interdisciplinary. These characteristics guide the development of a comprehensive evaluation framework. III. Research Questions Building on the theoretical framework, this study addresses the following questions: How can we evaluate the quality of university curriculum from the perspective of advanced knowledge? What is the overall quality of university curriculum in Chinese universities? What are the differences in the role of university curricula among different types of institutions? IV. Data and Methods Data for this study come from the 2021 Chinese College Students' Survey, covering 33 institutions and 118,249 participants. The study employs a theoretical framework above. Statistical tests and confirmatory factor analysis are used to validate the evaluation framework. V. Results The study reveals that overall curriculum quality in Chinese universities needs improvement, particularly in knowledge-specification, cognitive-challenging, research-oriented, and interdisciplinary. Poor curriculum quality caused students' superficial learning for exam purposes. Additionally, the study emphasizes the central role of university curriculum, with applied universities showing lower scores in various quality indicators compared to academic universities. VI. Research Applications This research has implications for policymakers, administrators, and faculty members. It highlights the need for a reevaluation of university curriculum and their unique contributions to talent development.

WEDNESDAY 3 JULY
Session 7.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

259. Utilizing the Ability to Identify Criteria (ATIC) to Select Personnel

Yolandi-Eloise Fontaine
Stellenbsoch University
François de Kock
University of Cape Town
Reinout E. de Vries
Vrije Universiteit Amsterdam
Eva Derous
Ghent University

For decades, researchers have been interested in determining how responding in hiring contexts relate to future employee performance. Some individuals perform well on testing criteria because they recognize the dimensions measured, and they know how to respond in a socially desirable way to create a positive impression. This ability to identify relevant evaluative criteria during selection procedures is referred to as ATIC, and is defined as a person's ability to perceive performance criteria correctly, when participating in evaluative situations. The aim of this research was to assess ATIC in a novel way and to determine if ATIC scores relate to performance. The sample consisted of university students (N = 178; 137 women and 41 men) aged between 21 and 53 years (M = 24.34; SD = 4.7). Data were collected in a sequential fashion, at three different points in time: ATIC and Social Desirability (Time 1 online); cognitive ability (Time 2, two weeks later and in a group-setting), Grade Point Averages (GPA; Time 3, eight weeks later, using official university records). To test the hypotheses a series of biserial correlations using SPSS v. 25 were conducted. The results showed that ATIC scores significantly related to cognitive ability ($r = .24$, $p < .001$) and GPAs ($r = .27$, $p < .001$); but not to social desirability ($r = .07$, $p = .34$). In line with other findings, measures of ATIC could be included during selections as a predictor of on-the-job performance. Future research and practical implications are discussed.

WEDNESDAY 3 JULY
Session 7.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

453. Using Critical Quantitative Methodology and MIMIC Modeling for Justice-Oriented and Antiracist Measurement

Matthew Diemer
University of Michigan
Michael Frisby
Georgia State University/USA
Aixa Marchand
University of Illinois

This presentation applies Critical Quantitative (CQ) methodology and MIMIC modeling to demonstrate their advantages for justice-oriented and antiracist measurement. White supremacy and bias permeate many of the measures we use – and so this talk aims to provide a broader methodological perspective (CQ; AUTHORS, 2023) and a specific methodological approach (MIMICs) to help identify, attenuate, and/or eliminate biased and/or racist items in commonly used measures. To achieve these aims, this presentation applies the CQ perspective and MIMIC modeling to a measure designed to capture the discrimination that Black people face, the Everyday Discrimination Scale (Sternthal, Slopens & Williams, 2011), to illustrate that two of the scale's five items under-measure racism for Black respondents. Data and code are available via an OSF repository. In the first step of MIMIC modeling, a CFA model was applied to 3,660 participants, aged 32-42, who identified as Black (N = 863, 23.57%) or as white (N = 2797, 76.42%). These data came from AddHealth, Wave V (Harris, 2013). The CFA model fit well and each item loading was strong and statistically significant. In the second step, an exogeneous covariate predicted the Everyday Discrimination latent variable. Black respondents had a significantly higher (β = .09) latent mean than white respondents. This path adjusts the latent means to be similar, before testing for item bias in step three. In the third step, two items were regressed onto the covariate. Black respondents reported poorer service in restaurants/stores (B = .20 $p < .001$) and that people more frequently act afraid of them (B = .11, $p < .001$), indicating DIF. The direction of effect indicates these items under-measure everyday discrimination for Black respondents, surprisingly. In short, this application of a broader CQ perspective and the specific MIMIC approach shows potential for anti-racist and/or justice-oriented measurement.

WEDNESDAY 3 JULY
Session 7.5
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

407. Executive functions in blind and deaf children: a Tablet-based assessment

Camila Martínez, Victoria Espinoza, Ricardo Rosas Pontifical Catholic
University of Chile

Executive functions are of the highest interest because of their relationship with learning, in general, and with reading and maths in particular. We hereby present the development of the YellowRed App, a Tablet-based test to assess Executive Functions (EF) in deaf and blind children. YellowRed was developed to assess the three main EF components (working memory, inhibition and flexibility) in typically developing children, as well as children with special educational needs, from 6 to 12 years of age. Additionally, as EF impairments could be associated with learning problems in deaf or blind children, two YellowRed versions adapted to these populations were developed. We present the results of a study performed in blind and deaf children by using the adapted YellowRed App, comparing their results to the typically developing children norm obtained in Chile. Results show only mild difficulties in executive functions development. Both blind and deaf children showed descended results at the working memory tests. On the other hand, for most inhibition (blind and deaf children) and cognitive flexibility test (only deaf children), their performance is close to the mean performance of typically developing children.The results support the usefulness of assessing EF in deaf and blind children by a tool that resembles that one for typically developing children, thus allowing for a direct comparison between different developmental circumstances.

WEDNESDAY 3 JULY
Session 7.5
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

688. Psychometric properties of Cattell's Fluid Intelligence Test (CFT 20-R) in Lithuanian sample of 8–17 year olds

Sigita Girdzijauskiene, Dovile Butkiene, Grazina Gintiliene
Vilnius University

In the context of school, the assessment of intelligence plays an important role in diagnosing special needs, e.g. special learning disabilities, giftedness. It is essential to develop an adequate test to assess the intellectual abilities of those pupils whose verbal representation is limited. CFT 20-R, Cattell's Fluid Intelligence Test (Wei , 2006) is language-free test which measures fluid intelligence. This is an ability to recognize figural relationships and formal-logical thought problems with varying degrees of complexity and process them within a certain amount of time. This test does not rely on language knowledge or cultural background, and so is particularly useful where fair assessment of children and adolescents from different linguistic and cultural backgrounds is required. The CFT 20-R is published in 12 countries, mostly with country-specific standardization. The aim of this study was to examine psychometric properties of the CFT 20-R in Lithuanian sample of children and adolescents. A sample of 2 872 children and adolescents (1 429 boys and 1 443 girls), from 2nd to 11th grade, were administered the full-length CFT 20-R with extended time limits. Sample size for test-retest reliability was 88 children. 296 children participated in validation study using Berlin Structure of Intelligence Test for Youth: Assessment of Talent and Giftedness (BIS-HB; Jäger et al., 2006). Split-half reliability for age groups ranged from 0.86 to 0.93 (Total score). Retest reliability for whole sample ranged from 0.69 (Part 1) to 0.76 (Total score). The validity of the CFT 20-R was supported by strong correlation between the CFT 20-R IQ and BIS-HB Reasoning (r=0.68), General intelligence (r=0.58), Figural abilities (r=0.56) and weak correlation between the CFT 20-R IQ and BIS-HB Creativity (r=0.27). To conclude, the CFT 20-R seems to be a reliable and valid instrument to assess fluid intelligence among Lithuanian children and adolescents.

WEDNESDAY 3 JULY
Session 7.5
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

692. **Validation of the IDS-15 among Italian Adolescents**

Adriana Lis, Silvia Salcuni, Giulia Bassi, Elisa Mancinelli, Vinay Jagdish Sukhija
University of Padova

Adolescents increasingly use the Internet for communication, education, entertainment, and other purposes in varying degrees. Given their vulnerable age, they may be prone to Internet addiction. Much literature has stressed the importance of validation of psychometric tools assessing Internet Addiction (IA). One of the newest proposed measures is the Internet Disorder Scale (IDS-15). The IDS-15 assesses the severity of IA in four distinct IA-related domains: (a) escapism and dysfunctional emotional coping, (b) withdrawal symptoms, (c) impairments and dysfunctional self-regulation, and (d) dysfunctional Internet related self-control. The scale assesses the impact of its effects by focusing upon users' online leisure activity (i.e., excluding academic and/or occupational Internet use) from any device with Internet access over the past year. This study aimed at investigating the psychometric properties of the Italian version of the IDS-15 by examining its construct and concurrent validity. 471 adolescents participants (Mage = 24.72 years, SD = 8.66; 256 males) were recruited from Italian secondary schools. The confirmatory factor analyses to determine the dimensional structure of the scale and invariance across genders will be discussed. Concurrent validity will be examined using wellbeing variables. Moreover, the reliability and validity of the scale will also be assessed. Keywords: internet-addiction, validation, adolescence

WEDNESDAY 3 JULY
Session 7.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

653. Generative AI and Large Language Models: Applications and Research Directions in Psychological Sciences

Hudson Golino
University of Virginia

The current symposium will discuss applications and research directions in the field of generative artificial intelligence and large language models in psychological research. The presentations will show how generative artificial intelligence can be used to 1) identify emotion dynamics in videos using zero-shot image classification and dynamic exploratory graph analysis, 2) study the psychometric quality of LLM-generated items using exploratory graph analysis and related techniques, 3) how to use LLMs to analyze data from students' evaluation of teaching, and 4) how generative AI can be used to improve quantitative psychology. Our symposium will provide a number of different applications and research directions combining generative artificial intelligence, network psychometrics, emotion research, quantitative psychology, and educational applications.

Discussant name: Hudson
Discussant surname: Golino
Discussant affiliation: University of Virginia

WEDNESDAY 3 JULY
Session 7.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

727. Decoding Emotions: Facial Expression Recognition with Transformer Models using the transforEmotion Package in R (Artificial Intelligence in testing, psychological assessment and survey research)

Aleksandar Tomasevic
Department of Sociology, University of Novi Sad
Hudson Golino
Department of Psychology, University of Virginia
Alexander Christensen
Department of Psychology and Human Development, Peabody College, Vanderbilt University

In this work, we introduce a new R package called `transforEmotion` for facial expression recognition of emotions (FER) using transformer-based machine learning models. Traditional machine learning approaches for emotion detection from images and videos require large amounts of labeled training data and are typically limited to detecting only basic emotions. However, zero-shot learning models based on transformer architecture, such as OpenAI's CLIP, can detect emotions from images and videos without the need for labeled training data. We provide a step-by-step workflow for leveraging the power of transformer-based models for emotion detection from images and videos using the `transforEmotion` package. This workflow is free, open-source, and does not require commercial APIs, access to GPUs, or knowledge of Python. Our goal is to empower the research community to utilize the capabilities of generative artificial intelligence to gain new insights into the dynamics of human emotion expression based on abundant data from social media and other sources.

WEDNESDAY 3 JULY
Session 7.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

65. Monticello Simulations: How Generative AI change how we do simulations in quantitative psychology (Artificial Intelligence in testing, psychological assessment and survey research)

Hudson Golino
University of Virginia

This presentation will discuss how generative artificial intelligence and large language transformer models can be used for implementing simulations in quantitative psychology. I will introduce a new methodology termed "Monticello Simulations" that can be used to study the quality of new quantitative methods in psychology. Monticello simulations will be compared to the traditional statistical approach of Monticarlo Simulations. An example using Exploratory Graph Analysis will be presented, showing what Monticello simulations offer above and beyond Montecarlo simulations.

WEDNESDAY 3 JULY
Session 7.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

766. Assessing the Quality of AI-Generated Items: A Network Psychometric Approach (Artificial Intelligence in testing, psychological assessment and survey research)

Lara Russell-Lasalandra, Hudson Golino
University of Virginia

Large Language Models' increasing ubiquity and efficacy have empowered researchers to explore once implausible projects, such as automatically generating thousands of new items. The present research introduces a new approach to assessing the quality of AI-generated items using the exploratory graph analysis framework without collecting data from humans.

WEDNESDAY 3 JULY
Session 7.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

658. Enhancing Student Evaluations of Teaching with Large Language Models: Insights from Active Learning Pedagogy (Artificial Intelligence in testing, psychological assessment and survey research)

Mariana Teles
University of Virginia

This presentation explores the innovative application of Large Language Models (LLMs) in analyzing student evaluations, focusing on their response to different teaching methodologies. Utilizing Facebook's BART Large Language Model, we conducted a zero-shot classification analysis on open-ended student feedback from two iterations of an Introduction to Cognition course. The first was taught in 2022 through a traditional lecture-based format, and the second, in Spring 2023, employed a student-centered active learning approach. Our analysis revolved around four key categories: "Better Learning Experience", "Learned More", "Engaged More", and "More Excited About the Content". These categories were chosen to capture the essence of students' experiences and their perception of the teaching methodologies. Zero-shot classification enabled us to analyze these open-ended responses without the need for pre-labeled training data, showcasing the versatility and depth of LLMs in educational research. The results revealed a significant enhancement in the active learning course across three categories: better learning experience, higher engagement, and more effective learning. These were quantified through Z-scores for a standardized comparison. Interestingly, the excitement about the course content did not significantly differ between the two teaching methods. This study not only demonstrates the practical application of LLMs in educational settings but also provides valuable insights into the effectiveness of student-centered active learning strategies. By leveraging the advanced capabilities of generative AI, we offer a novel perspective on evaluating and improving teaching methodologies.

WEDNESDAY 3 JULY
Session 7.7
Topic: Translation of tests, psychological assessment instruments and survey questionnaire / Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Identifying biases by qualitative or quantitative methods

209. Unveiling Cultural Biases: Exploring the Efficacy of Ipsative Methods in Recruitment Tools in Indonesia

Adji Pradana, Maharani Syahratu Kertapati, R. Brahma Aditya
Daya Dimensi Indonesia/Indonesia

Developing a recruitment tool free from biases related to desirability and cultural influence presents a significant challenge, and the ipsative method is one solution to deal with this challenge. However, ipsative measuring instruments in Indonesia are still underdeveloped and seldom used. This study aims to investigate the effectiveness of the ipsative method in reducing the influence of desirability and culture. The research uses two assessment tools—personality and motivation tests—employing the ipsative method, where each item is paired based on previously calculated levels of desirability during the pilot phase. The total number of participants in this research is 4.471 respondents, with detail 2781 for the personality test and 1690 for the motivation test. The findings indicate that despite employing the ipsative method, it fails to reduce desirability and cultural biases. In the personality assessment, the conscientiousness dimension consistently emerges as the primary choice compared to other dimensions. Similarly, in the motivation assessment, the dimension emphasizing social impact consistently takes precedence, while obtaining free time consistently ranks lower than other dimensions. The outcomes of both personality and motivation tests show disparities in response choices concerning dimensions perceived as advantageous for job applications in Indonesia.

WEDNESDAY 3 JULY
Session 7.7
Topic: Translation of tests, psychological assessment instruments and survey questionnaire / Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Identifying biases by qualitative or quantitative methods

232. Differential Prediction in an African Context: Analysing Personality and Sources of Bias in Selection

Pakeezah Rajab
JVR Africa Group/South Africa
Andrew Morris, Jani Wiggett
JVR

Research exploring bias provides a foundational understanding of differential prediction (subgroup differences in test-criterion regression equations) (SIOP, 2018). In the South African context, however, despite ethical and legal concerns associated with adverse impact, ethnic predictive bias differences on personality tests have rarely been explored. This study used stratified sampling, with 301 employees (White = 63.1%; Black = 36.9%) representing multiple industries and job levels completing the Hogan Personality Inventory (HPI), a Five Factor Model-based employment-oriented measure (Hogan & Hogan, 2010). Managers also rated these employees' performance using the Individual Work Performance Review (Van Lill & Taylor, 2022), across in-role, extra-role, adaptive, leadership and counterproductive performance areas, which was also combined into a general performance factor. The objective was to determine if ethnic prediction bias exists in the HPI and if so, whether performance is over- or under-predicted for different ethnicities. Numerous statistical techniques were therefore employed - dominance analysis identified that Inquisitive and Ambition HPI scales consistently predicted extra-role, adaptive, leadership and general performance. HPI and performance correlations were small but significant (.10 to .26). Effect sizes were small to moderate (-.05 to .46) across ethnicities. Mann-Whitney U tests only revealed differences for Inquisitive (W = 10561, p = .037). Step-down hierarchical regression indicated minimal differences in intercepts for personality scales depending on the criteria and only one significant slope difference indicative of potential over-prediction for Blacks. Our findings emphasise the need for appropriate tools in cross-cultural assessment and research, fostering equitable and inclusive decision-making processes, ensuring that talent is recognised and nurtured regardless of ethnic backgrounds.

WEDNESDAY 3 JULY
Session 7.7
Topic: Translation of tests, psychological assessment instruments and survey questionnaire / Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Identifying biases by qualitative or quantitative methods

306. The discrepancy between manifest response on a Likert-type scale and the most fixated response option as an indicator of social-desirable responding

Stanislav Ježek, Martin Burget
Faculty of Social Sciences, Masaryk University
Martin Jakubek, Monika Krafčíková
Faculty of Arts, Comenius University Bratislava

Eye-tracking studies have been showing gaze fixations are a useful source of information about cognitive processing of questionnaire items. We analyzed an existing dataset (Jakubek & Krafčíková, 2016) comparing observed responses with responses derived from eye-tracking (ET) data. We derived 2 ET responses: the most fixated option (ET response) and a weighted mean of fixated options (weighted ET response). The dataset comprised responses to Slovak version of NEO-FFI with 5-point Likert-type response scales administered to 50 university students. The inventory was administered twice in random order; in one condition the participants were instructed to respond honestly and in the other fake good. In planned analyses of the honest-condition responses we found a high level of agreement between the observed responses and the ET responses; the overall agreement was 86%. The observed scale scores correlated with the scores based on weighted ET responses from .92 to .97. The weighted ET scores have slightly smaller variances than manifest scores and their popularities are slightly and systematically different. The fit and parameters of unidimensional CFA models did not systematically differ between observed and weighted ET responses. Additionally, post-hoc analyses suggested that the discrepancies between observed and weighted ET responses were in the direction of social desirability and that the discrepancies correlate highly with the popularity of items in the fake-good condition. We discuss whether the discrepancies represent social-desirable editing during the response phase of responding as conceptualized by Tourangeau et al. (2000) or are an artefact of boundedness of the response scale. These findings are based on a limited sample not allowing us to fit multidimensional models. Data were originally collected for other purposes and contain limited ET information. The agreement of observed and ET responses is somewhat inflated by the use of mouse for responding.

WEDNESDAY 3 JULY
Session 7.7
Topic: Translation of tests, psychological assessment instruments and survey questionnaire / Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Identifying biases by qualitative or quantitative methods

700. A Methodological Approach to Evaluating and Selecting Overclaiming Items

Felipe Valentini, Leonardo Botinhon de Campos, Leticia Silva de Souza, Nelson Hauck, Ricardo Primi, Sunshine Martins, Anny Gabrielle de Almeida
University São Francisco, Brazil

The exaggerated self-report of positive psychological characteristics can distort scores in evaluations. This phenomenon manifests itself in both personality traits (such as kindness and joy) and cognitive abilities (for example, claiming proficiency in a language or task). Specifically, cognitive self-enhancement can due to the overclaiming bias, which can artificially inflate an individual's skill results. To measure overclaiming, questionnaires containing true and foil items are used. Scores are assessed according to three indicators: knowledge (endorsement of true items), false alarms (endorsement of foil items), and bias (indiscriminate endorsement of both types of items). However, a consolidated methodology to assess the quality of these items in questionnaires is still lacking. This study aims to present a method to analyze the quality and select appropriate items to measure overclaiming. We propose to evaluate the correlations between the scores of each item with the indicators of knowledge, false alarm, and bias. It is expected that good true items will correlate positively with the knowledge indicator and not with false alarm. Ideal false items should correlate positively with the indicators of false alarm and bias. However, false items that also correlate with knowledge can be considered "trick" items. We illustrate the method using a real database, where an overclaiming questionnaire was applied to 400 university students under two different experimental conditions: an honest response and another as if they were in a job selection process. It was observed that good true items showed positive correlations with the knowledge indicator. However, one foil item showed positive correlation with knowledge, suggesting a "trick" content. This study contributes to the construction of more effective instruments to measure overclaiming.

WEDNESDAY 3 JULY
Session 7.8
Topic: Validity theory in testing, psychological assessment and survey research

48. Psychometric Parsimonious Parameterization for Evidential Accuracy and Precision

Joshua Chiroma Gandi, Hauwa Mary Aigboje
Nigerian Defence Academy, Kaduna, Nigeria

Psychometrics is the science of psychological measurement which mostly focuses on the nearest approximation of phenomenon under consideration. However, being inappropriately parsimonious in the process constitutes failure to adequately model phenomena while (on the other hand) excessive modeling of a phenomenon amounts to over-parameterization. This challenge requires deliberate strategies to ensure evidential accuracy and precision. Hence the study, which sets to elucidate parsimonious parameterization for evidential accuracy and precision, adopted systematic review and meta-analysis based on explanatory design. 112 candidate studies were initially generated and, thereafter, 15 of these studies which satisfied inclusion-exclusion criteria are retained. Statistical techniques employed are the two one-sided tests (TOST) for equivalence test which computes effect size to facilitate determining inference(s), Neyman-Pearson analysis which pre-specifies type 1 error rate, and I-squared (I2) statistic which checks for heterogeneity. The summary of results (0.9 and 1.1 [H0: P1/P2 < 0.9 or P1/P2 > 1.1 versus H1: 0.9 < P1/P2 < 1.1]) rejected effect sizes larger than the equivalence bounds which pre-specified type 1 error rate and then showed the true effects as extreme as the equivalence bounds. By indicating a balance between parsimonious parameterization versus evidential accuracy and precision, the study supports a key assumption which corroborates that the former facilitates ensuring significant validity of the latter. Since the originating equation captures important features of evidential accuracy and precision that parsimonious parameterization purports to represent, any resulting inference applies equally to overall psychometric properties and not only to itself. Therefore, a carefully handled parsimonious parameterization is considered a significant match for determining psychometric soundness which lends credence to measurement quality.

WEDNESDAY 3 JULY
Session 7.8
Topic: Validity theory in testing, psychological assessment and survey research

111. Advancing Linguistically and Culturally Fair and Community-Relevant Assessments

Pōhai Kūkea Shultz
University of Hawaii/United States
Kerry Englert
Seneca Consulting/United States

The State of Hawai'i in the United States encompasses a particularly diverse population. Most notably, the Native Hawaiian community has survived attempts at marginalization and cultural suppression. However, over the last 40 years, the community has advocated for inclusion in the state's education system through the establishment of Hawaiian language immersion (Kaiapuni) schools. The community has also tirelessly advocated for fairer tests for their students. Thus, the Kaiapuni Assessment of Education Outcomes (KĀʻEO) has replaced English language tests for federal accountability. The KĀʻEO represents a culturally and linguistically fair model for Kaiapuni students and codifies the value of Hawaiian language and culture in the state's education and assessment system. Since the KĀʻEO's inception, we have focused on cultural and community validity as foundational values for the assessment (Shultz & Englert, 2023; Solano-Flores, 2001). Cultural validity provides a lens through which to view and deeply reflect on cultural and linguistic issues related to content by examining the values, knowledge, and skills of students and the community. Community validity seeks to engage diverse stakeholder groups to ensure their priorities and concerns for their students' education are measured in the assessment. In the proposed session, we will elaborate on the definition and importance of cultural and community validity for the KĀʻEO and how these concepts can be leveraged in other assessments. For example, we will provide findings from a community validity survey with teachers to better understand the alignment of KĀʻEO results with their data use priorities. Even though the KĀʻEO is a smaller statewide assessment, we assert it is not only possible to build a foundation of cultural and community validity in large-scale assessments, but necessary if we truly aim to develop culturally and linguistically fair and community-relevant assessments.

WEDNESDAY 3 JULY
Session 7.8
Topic: Validity theory in testing, psychological assessment and survey research

409. Reduction of Faking Using a Forced-Choice Format: Is It Pancultural?

HyeSun Lee, Weldon Smith
California State University Channel Islands

Forced-choice format tests have been proposed as an alternative to Likert-scale measures for personnel selection due to robustness to faking and response styles. This study conducted a comparative analysis of the extent of faking in Likert-scale and forced-choice five-factor personality tests, focusing on South Korea and the United States. Additionally, the effectiveness of the forced-choice format in reducing faking was explored in both countries. Data were collected from 396 incumbents participating in both honest and applicant conditions (N_South Korea = 179, N_US = 217). Cohen's d values for within-subject designs (ds_within) were employed to assess the magnitudes of faking in each format and country. In both countries, the degrees of faking in the Likert-scale were larger than those from the forced-choice format, and the magnitudes of faking across five personality traits were larger in South Korea by from 0.07 to 0.12 in ds_within. The forced-choice format demonstrated a successful reduction in faking for both countries, as evidenced by an average decrease of 0.06 in ds_within for each. However, the patterns of faking in the forced-choice format exhibited variations between the two countries. In South Korea, there was an increase in faking degrees for Openness and Conscientiousness, while degrees of faking for Extraversion and Agreeableness substantially decreased. The presentation discusses potential factors contributing to trait-specific faking in the forced-choice format, taking into account cultural influences on the perception of personality traits and score estimation in Thurstonian IRT models. Finally, it explores perceptions of forced-choice formats in comparison to Likert-scale formats, based on qualitative data collected from 432 Asian-American respondents.

WEDNESDAY 3 JULY
Session 7.8
Topic: Validity theory in testing, psychological assessment and survey research

537. Factor Analysis under multimodal latent distributions: A simulation study

Oscar Lecuona
Universidad Complutense de Madrid
Victor Ciudad
Universitat de València
Guido Corradi
Universidad Villanueva

Conceptual framework: Factor analysis is one of the most popular techniques to estimate latent variables in social sciences, both in their exploratory and confirmatory branches. Among others, the factor model assumes that latent variables are unimodal and normally distributed. This assumption can be challenging when working with several populations and cultural groups due to potential multimodal distributions, which impedes fairness and generalizability of research findings. Objectives: To investigate the impact of multimodal latent variables on factor analyses, specifically examining the potential bias in parameter estimation. Methodology: This simulation study approaches this question by applying exploratory and confirmatory factor analyses with multimodal latent distributions. More concretely, we created several latent variables sorted in the severity of their multimodality, both in the number of modes to their distance between them. Results: As multimodality increases, factor weights exhibit a heightened bias towards positive values, which may be an aberrant cue for researchers. Fit indices and parallel analyses showed also biases with specific patterns as multimodality gets more pronounced. Only factor scores showed up as reliable descriptors of the latent multimodality. Implications: Latent multimodal distributions can be a relevant challenge in latent variable models, which underscore the importance of careful consideration and interpretation by researchers. We suggest more research on factor scores as potential reliable indicators of multimodal latent variables, while also the limitations and applications of these findings.

WEDNESDAY 3 JULY
Session 7.8
Topic: Validity theory in testing, psychological assessment and survey research

684. Gender DIF and gender differences in civic outcomes over time

Yuan-Ling Liaw
IEA Hamburg

Gender differences in reading and mathematics are acknowledged in international large-scale assessments (ILSAs) across countries. However, limited exploration exists on gender disparities in civic knowledge, attitudes, and engagement. This study addresses the gap by investigating factors contributing to gender differences in young people's civic outcomes over time. Using data from the IEA's International Civic and Citizenship Education Study (ICCS) in 2009, 2016, and 2022, this research aims to illuminate gender differences in civic outcomes across these years. In the recent ICCS 2022, with 22 countries and 2 benchmarking entities, girls consistently demonstrated higher global civic knowledge, sustaining a pattern observed in three cycles. ICCS 2022 also highlighted global support for gender equality, with female students expressing stronger endorsement than male peers. Despite girls potentially outperforming boys, indicating advantages in civic knowledge and engagement, it's essential to recognize that the ICCS evaluates not only students' civics and citizenship knowledge but also includes reading comprehension. The observed gender difference in ICCS performance may be linked to better reading comprehension skills among girls. To further explore gender differences, the study investigates whether they can be linked to gender-specific differential item functioning (DIF). Findings reveal associations between DIF and a country's economic development, geographic location, democracy index, linguistic background, and socio-cultural characteristics like religion. This exploration enhances our understanding of gender differences in civic knowledge across countries and time, providing insights for educational policies and practices.

WEDNESDAY 3 JULY
Session 7.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

121. Development and Validation of an Integrative Employee Well-being Scale

Clara Y.W. To, Evan, F.H. Choi, Jerry Chan, Hodar Lam
DIOP, The Hong Kong Psychological Society

Wellbeing is a multi-dimensional construct of individual positive functioning, with indicators such as physical health, emotional wellbeing, social connections, and life satisfaction. Organizations are increasingly realizing the significance of workplace wellbeing for their employees' job satisfaction, productivity, and overall business performance. There is a growing need for a more employee-oriented approach to workplace wellbeing, which emphasizes diversity and inclusion. Yet assessments of workplace wellbeing seem to be fragmented and lack a comprehensive approach. Thus, comparison and integration of prior empirical work are often challenging. It is also important to note that the definition of wellbeing may vary depending on cultural and contextual factors. To address these issues, we conducted two cross-sectional survey studies between 2020 and 2023. In Study 1, we developed a 17-item inventory to collect responses from 361 adult samples in different locations, including Hong Kong, Mainland China, the UK, and Europe. Confirmatory factor analysis results revealed a four-factor model of wellbeing, which are: physical, psychological, community, and meaning in life respectively. In Study 2, the original survey was expanded from 17 to 53 items to cover a wider range of wellbeing facets. Apart from the four dimensions identified in Study 1, we developed items to measure three additional wellbeing dimensions: financial and spiritual wellbeing, as well as the absence of negative emotions. A total of 305 adult samples from Greater China were analyzed, and our data revealed a seven-factor model of wellbeing through confirmatory factor analysis with a satisfactory model fit. Our findings inform the possibility for integrative and customized approaches to workplace wellbeing assessments. Accurate assessments can foster organizational and strategic awareness of employees' different wellbeing aspects and targeted interventions and policies with cultural relevance.

WEDNESDAY 3 JULY
Session 7.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

429. Sociocultural Approaches to Assessing Work-Aligned Engineering Competencies

Maria Elena Oliveri
Buros Center for Testing
Kerrie Douglas
Purdue University

This presentation illustrates the application of a Sociocultural Evidence-Centered Design (ECD) model crafted for designing and validating multidimensional assessments targeting engineering competencies. It addresses the increasing demand for engineers who can integrate content knowledge with practical skills in response to evolving workplace competencies. The model guides developers and researchers from multidisciplinary teams in engineering education, demonstrating why and how to assess multidimensional engineering competencies to meet the evolving needs of the engineering profession. Key components involve utilizing a Q matrix to integrate Content Knowledge and Engineering Practices, emphasizing assessments that surpass traditional evaluations, focusing on engineers' seamless integration of content knowledge with practices and technical skills across domains. Challenges in creating high-quality engineering assessments will be discussed, including enabling diverse learners to communicate effectively and meeting the global demand for engineers with profound understanding of complex issues. The ECD model recognizes the significance of multidisciplinary integration, advocating for a holistic learning approach to enhance the educational experience and prepare engineers for the multifaceted demands of the contemporary engineering landscape. The presentation will illustrate the use of a Sociocultural ECD model. The presentation will be useful for developers and researchers to develop and validate assessments that effectively measure multidimensional engineering competencies. By addressing challenges and embracing diverse learner needs, this presentation demonstrates the application of a method to develop flexible, authentic, and comprehensive assessments for complex and multidimensional engineering competencies.

WEDNESDAY 3 JULY
Session 7.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

495. Examining Methodologies for Contextual Information in Large-Scale Assessments

Priyanka Sharma
India

Large scale studies of educational achievement conducted at national and international level simultaneously employ cognitive instruments and questionnaire to gather evidence of outcome variables and contextual information that can explain variation in the outcome variables. Cognitive instruments generally adhere to established measurement principles with a structured assessment and analytical framework. On the other hand, a noticeable disparity exists in both structure and methodology of contextual instruments or surveys, presenting substantial challenges to the extraction of meaningful insights for specific educational settings. This persists despite the fact that contextual questionnaire has a significant role in overall study design and has evolved exponentially over time to incorporate students' attitudes, dispositions, and non-cognitive outcome measures as compared to the conventional social, cultural, and economic factors. This paper aims to examine technical considerations and associated challenges with contextual questionnaires, with the objective of enhancing their efficacy and utility for informing policy measures. The analysis delves into the conceptual frameworks of national and international assessments, analysing their constructs, variables, and indices. Additionally, a comparative analysis of the design and analytical approaches has been employed. Drawing upon these analyses, the paper proposes feasible options to augment the effectiveness of contextual instruments, enhancing their efficacy in addressing policy inquiries and enabling evidence based policy decisions.

WEDNESDAY 3 JULY
Session 7.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

773. Receptivity to Instructional Feedback: A Cross-cultural Validation Study

Lopera Oquendo, Anastasiya Lipnevich
City University of New York
Samuel León Parra
Universidad de Jaén

Instructional feedback is widely acknowledged as crucial for motivating students and achieving academic goals. However, its effectiveness depends on the student's disposition and ability to use it. Lipnevich and Smith (2016, 2022) proposed a model of student-feedback interaction, emphasizing student receptivity to feedback as a key learner characteristic. The concept of feedback receptivity posits that differences exist in how individuals are willing or ready to accept feedback. While these differences may be situational and context-dependent, there appears to be a trait-like feedback receptivity that persists across diverse scenarios, The Receptivity of Instructional Feedback scale (RIF), a well-validated instrument, comprises 24 items, rated on a 5-point Likert scale, distributed across four factors. RIF has been translated, adapted, and administered across diverse educational contexts and samples. This study involved 4,635 middle, high school, college, and master students from Spain, the United States, New Zealand, Singapore, Turkey, and Brazil. Using correlational analysis, Exploratory Structural Equation Modeling (ESEM), Network Analysis, and Differential Item Functioning (DIF) through MIMIC modeling, we examined its internal, cross-cultural, and discriminant validity. Findings indicate: a) the hypothesized 4-factor model was well-supported across different samples; b) a bifactor-CFA model revealed a general receptivity factor alongside the four specific factors; c) SEM and DIF analysis supported the measurement equivalence invariance; and d) the RIF predicted relationships with external variables (e.g., personality and grades), supporting discriminant validity. Overall, results underscore the significance of validation efforts in enhancing psychometrical adaptability across diverse educational contexts. Instruments like RIF can serve as valuable and well-validated tools for comprehending students' responses to instructional practices and interventions.

WEDNESDAY 3 JULY
Session 7.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

816. Comparison of two methods to obtain the dimensionality of an instrument: factorial analysis Vs. Rasch analysis

Angélica Garzón Umerenkova
Fundación Universitaria Konrad Lorenz - Facultad de Psicología (Bogotá-Colombia)
Jesús de la Fuente Arias
Universidad de Navarra - Facultad de Educación y Psicología (España)
Érika Andrea Malpica
Secretaría de Educación de Bogotá (Colombia)

In the Rasch model, the dimensionality is established through Principal Component Analysis (PCA), which is different and more parsimonious than the one performed on TCT with Factorial Analysis. The aim is to compare these two methods using the dimensionality analysis of a test. The instrument Procrastination Assessment Scale-Students (Solomon & Rothblum, 1984) is the most used in academic procrastination research with children and teenagers. We analyzed the second part of the PASS (25 items), which inquiries about the cognitive-behavioral reasons to procrastinate. For this, different factorial solutions are reported by cultures and countries, which generates confusion and difficulties for transcultural comparison. Sample: 451 students from a public school in Bogotá (Colombia). Two hundred two males (44.8%) and 249 females (55.2%), and ages from 10 to 18 (M =13.99; SD = 1.91). Method: PCA and Varimax factorial rotation were used for the Exploratory Factor Analysis (EFA). Three factors that explain 44.3% of the variance and exclude four items from the test were obtained. The internal consistency for the three factors is between .71 and .82. For the Rasch analysis, PCA is performed, considering that the first residual does not surpass 3 own values. Only one factor was obtained, which explains 59% of the variance, and all items are preserved. The item's reliability is .99, and the persons´ have a separation index of 2-79 (.89). Conclusions: The results indicate that compared to the Rasch analysis, EFA is a method that tends to multiply a test´s factors. Likewise, in this case, by means of the Rasch analysis, all items could be preserved, and a good person´s reliability indicator was obtained, which cannot be estimated using classical test theory (CTT). Implications: The Rasch analysis may be a promising method for adapting multidimensional CTT tests. Limitations and reach are discussed. Funding: Effect of the Degree of Regulation (personal and contextual) on Competence for Psychological Well-being, Emotional Health, and Flourishing, in Different Psychological Contexts. The project led by Jesús de la Fuente Arias (PR), with file number PID2022-136466NB-I00 (Universidad de Navarra), has been financed by the 2022 call for "Proof of Concept" Projects of The Spanish State Research Agency (AEI). It is co-financed by the European Regional Development Fund (FEDER).

WEDNESDAY 3 JULY
Session 8.1
Topic: Innovations in test development

324. How modern psychometrics can change intelligence testing in children and adolescents

Tatjana Kanonire, Ekaterina Orel
HSE University
Tatiana Logvinenko
Sirius University of Science and Technology
Alena Kulikova
HSE University

Despite the long history of intelligence testing, or perhaps because of it, widely used tests such as the WISC or Stanford-Binet are limited in their application of modern psychometric approaches by their linear, paper-based format. However, the development of an entirely new modern intelligence test, that can be used to comparably measure children and adolescents alike with the recent psychometric advances can overcome the mentioned limitations of the existing tests. In this study we analyse the use of computer-based testing and innovative psychometric technologies such as technology-enhanced items, the utilization of universal test design principles, the application of multidimensional IRT models, and the adoption of computerised adaptive testing. We discuss how these psychometric technologies can change not only the test format and its features but also our understanding of intelligence. The issues of construct and criterion validity, the incorporation of repeated measures, and the prediction of individual progress are discussed.

WEDNESDAY 3 JULY
Session 8.1
Topic: Innovations in test development

560. Unlocking Natural Language Processing: Predicting Item Difficulty in the Brazilian High School Exam (Enem) Without Pre-Testing

Alexandre Jaloto
National Institute for Educational Studies and Research Anísio Teixeira (Inep)
Alexandre José de Souza Peres
Federal University in Mato Grosso do Sul (UFMS)
Ana Carolina Zuanazzi
Airton Senna Institute (IAS)
Araê Cainã, Ricardo Primi
São Francisco University (USF)

The Brazilian High School Exam (Enem) is annually administered to over three million students, serving as a criterion for admission to higher education. With a minimum of two annual applications, test equating is carried out using pre-tested items. However, this approach entails high operational costs and risks to the process, given the high-stakes context of the exam. A viable alternative for pre-testing involves the application of Natural Language Processing (NLP) tools to determine item parameters in the Item Response Theory (IRT). Thus, this study aimed to assess the effectiveness of predicting the difficulty parameter of Enem items based on their textual characteristics. We employed a linear regression approach with machine learning to predict the difficulty parameter of Natural Sciences items. Predictor variables consisted of the mean item score across each of the 300 dimensions of trained word vectors (word embeddings). The sample of 600 items was divided into a training set (480) and a test set (120). The correlation between the b-parameter of the test set and the predicted value reached 0.50. Dimensions that positively contributed the most in the regression were associated with specialized vocabulary and specific contexts in the field of Natural Sciences (e.g., atom, hydroxides, capacitors), while those with the greatest negative contribution presented less specialized and more general vocabulary (e.g., mediated, swine, injuries). The results suggest that predicting item difficulty based on textual characteristics is a promising approach for high-impact tests like Enem, potentially reducing costs and risks associated with pre-tests by decreasing the need for subject responses to calibrate items. We recommend that future research explores other procedures for obtaining numerical item vectors, such as the use of BERT.

WEDNESDAY 3 JULY
Session 8.1
Topic: Innovations in test development

775. Automatic Item Generation for Large-scale Assessment Instruments: A Mexican Perspective

Citlalli Sanchez-Alvarez
Universidad Autónoma de Baja California/Mexico

Throughout most of the last century, educational assessment predominantly focused on paper-pencil based tests featuring multiple-choice response items. However, recent advancements have transformed the landscape of large-scale assessment instruments, exemplified by renowned assessments like PISA, TOEFL, SAT, and GRE. These modern instruments integrate insights from cognitive theories, psychometrics, and computer technology. Automatic Item Generation (AIG) emerges as a pivotal methodology, aiming to automate the item development process and generate a substantial quantity of items that are both conceptually and psychometrically equivalent. This paper explores the application of AIG to enhance large-scale assessments, acknowledging the existing challenges. Two fundamental approaches have been employed in AIG item development: strong theory and weak theory. In the former, a cognitive model is constructed to identify underlying cognitive processes and elements influencing item difficulty (Gitomer & Bennett, 2002). The latter involves selecting a parent item from a calibrated pool or developing a new one, with item developers relying on theoretical and practical expertise to identify non-influential characteristics (Drasgow, Luecht & Bennett, 2006; Gierl & Lai, 2012). This research focuses on the development of large-scale assessment instruments by Mexican researchers through AIG principles. A multi-stage development model is introduced, showcasing various item models designed to streamline the item development process. The study addresses the theoretical and conceptual framework, and implications of implementing AIG in the Mexican educational context. The findings contribute to the evolving field of large-scale assessment, shedding light on the innovative strategies employed by researchers in instrument development.

WEDNESDAY 3 JULY
Session 8.2 SYMPOSIUM
Topic: International assessment

214. Personality and Potential Across Languages and Cultures

Lauren Jeffery-Smith
Saville Assessment, United Kingdom

Organisations do not operate in a cultural vacuum. Technological advancements mean individuals can work globally so need to be able to collaborate with, manage and lead individuals from multiple different cultures. Even organisations only operating in one geography have individuals from different cultures working within them. Assessing personality cross-culturally raises considerations in terms of the generalisability of personality scales and the equivalence of instruments in different languages (McCrae, 2002). When developing assessments, particularly those involving complex constructs, the impact of potential cultural differences needs to be taken into consideration. By taking an ETIC approach, many psychometric test publishers aim to provide assessments which can work cross-culturally by developing based on international data, then comparing data by regions and countries to investigate equivalence. The first paper will discuss research into machine versus human translation in the context of adapting psychometric assessments for use in different languages. The potential benefits, considerations and drawbacks of using AI for translations will be discussed. The second paper will explore the application of a leadership potential assessment across different cultures. The assessment, which has been developed based on international data, will be compared in different regional samples to understand any differences in the level of expression of leadership career indicators. The final paper will discuss the simultaneous development of a personality assessment in different languages and how this is beneficial for the reliability of scales across different language versions.

WEDNESDAY 3 JULY
Session 8.2 SYMPOSIUM
Topic: International assessment

217. Assessing Leadership Potential Across Different Cultures (International assessment)

Lauren Jeffery-Smith, Rab MacIver
Saville Assessment, United Kingdom

While five factors of personality have been consistently found across cultures, the level of expression of these traits may not necessarily be consistent (McCrae & Costa, 2008), particularly in more diverse cultures e.g., individualistic versus collective (Hofstede, 1980). These five factors can be used in the identification of leadership potential with other variables including motivation. It is important to explore the expression of personality markers of potential across cultures. This paper explores the use of a framework for identifying leadership potential across different cultures. The framework studied includes a measure of overall leadership potential and three career area indicators – Professional, People and Pioneering leadership (identified from research into leadership effectiveness and personality correlates). This acknowledges that there is not one single way to be an effective leader and different people will be suited to different paths within an organisation. Algorithms were developed based on international samples comprising over 7,000 independent ratings of performance and potential across roles and organisations. The algorithms that resulted were then optimized using international data on over 18,000 individuals. This process reduced already small group differences on gender, age and ethnicity. To explore these algorithms across different cultures, scores for regional samples from Africa, Asia, Europe, Latin America, the Middle East and the USA were compared. Between the regional samples, no differences greater than 1 Sten were found. The largest differences were the Africa group being approximately .8 of a Sten higher than the Asia and Europe groups on the Professional indicator. This is consistent with previous research where Africa scored higher on Conscientiousness than other regions (Schmitt et al., 2007). As there are no notable differences between the regional groups on the measures, this supports the use of the framework globally.

WEDNESDAY 3 JULY
Session 8.2 SYMPOSIUM
Topic: International assessment

218. Multi-lingual Development of the Great 8 Success Factors (Translation of tests, psychological assessment instruments and survey questionnaire)

Rainer Kurz
HUCAMA

The objective of this paper is to share experiences gained in the development of the Great 8 Success Factors and 48 sub-ordinated constructs. Following an in-depth review of contemporary personality and competency assessment models NEO IPIP items were adopted from the multilingual source and augmented with items created to expand occupational coverage resulting in 758 items distributed across two questionnaires. English, German, and Danish versions were prepared by the authors (native seakers) and Swedish, Spanish, and French versions translated by local native speakers, and then back translated. 466 professionals and managers completed both assessments. Initial item analysis, PCA and scale constructions proceeded across all language versions. The average reliability of the 30 NEO IPIP constructs was .79 and .76 for the Personality Factors facets. The standardisation version of Personality Factors contained 240 items to measure personality 48 facets. Some items were refined in the light of the hierarchical model that emerged which pairs up DeYoung's (2015) Stability and Plasticity meta-factors with People and Task themes derived from the Ohio leadership studies in the 1960s. The model builds on the Universal Values circumplex of Schwarz (1992) and features four quadrants, 8 factors and 48 facets that are structurally aligned across predictor and criterion measures. Based on instrument usage a 'Professionals & Leaders' norm (N=1079) was compiled in early 2023 with a median internal consistency of .76 for the 48 personality facets. Completions in English (N=562), German (N=246), Swedish (N=198) and Danish (N=71) resulted in median reliabilities of .77, .75, .69 and .74 respectively. Biographical data largely accounted for language differences. Implications of this research are that NEO IPIP items can form a powerful starting point for instrument development and that simultaneous development in multiple languages is beneficial.

WEDNESDAY 3 JULY
Session 8.2 SYMPOSIUM
Topic: International assessment

402. Assessments and Machine vs Human Translation (Translation of tests, psychological assessment instruments and survey questionnaire)

Camille Stevenson
Saville Assessment

Advancements in artificial intelligence (AI) in recent years have seen machine translation become increasingly considered and tentatively implemented in the translation industry. AI-powered translation tools, such as neural networks, have surpassed all developments in machine translation (MT) of the past two decades (Yuan & Sharoff, 2020). However, we must consider the limitations of MT for assessment adaptation and how it can be appropriately utilised to drive progress in cross-lingual testing. This paper will discuss the benefits and drawbacks of AI in the adaptation of psychometric assessments, discussing ramifications in both the translation process and the product. Incorporating evidence from literary reviews, in combination with our own MT experiments, this paper argues that AI and MT cannot be used without the aid of human translation. During our trial, we noted the significant improvement in quality from MT, particularly for languages that were previously considered too challenging for such systems. We ran an initial trial with 'simple' platform text and found that the MT output contained fewer lexical and grammatical mistakes than previous MT technologies; at times, the machine translation was indistinguishable from human translation. However, the system struggled significantly when not given enough semantic context and misunderstood the meaning of more technical source text. Thus, we needed human translators to review the MT output. Whilst AI shows potential for enhancing cross-language adaptation, ethical considerations must be kept in mind. Translators hold valid concerns about what the use of MT means for them; there is a need for further training to articulate the benefits of MT and post-editing for translators. Moreover, MT is not yet advanced enough to be used independently. If incorporated correctly, AI can lead to a host of new opportunities in the translation of psychometric assessments – but, for now, only in collaboration with human translators.

WEDNESDAY 3 JULY
Session 8.3 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

657. Test Validity 2.0: Mathematics, Quantum Information Theory, Movies, Music and Jokes

Hudson Golino
University of Virginia

The current symposium will discuss test validity using new lenses, from mathematical theories to quantum information theory. The goal is to go above and beyond the traditional psychological perspectives of validity and make a broader, provocative discussion on out-of-the-box ideas. The symposium will have only two presentations and (hopefully) sufficient time for a back-and-forth between the presenters and the attendees, with lots of good humor and explicit and implicit references to famous and not-so-famous movies and music. If you are usually bored at conference presentations (just like me), then this symposium will (hopefully) function as fresh air despite the high temperatures in Granada in July (pending a good working air conditioning system). Nobody would miss a symposium with a title combining quantum information theory, mathematics, movies, music, and jokes, right?

WEDNESDAY 3 JULY
Session 8.3 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

802. Big Questions with Even Bigger Psychometricians: The Construct of Egomania Rageosis and The Theory of Everything (Translation of tests, psychological assessment instruments and survey questionnaire)

Bruno Zumbo
University of British Columbia

In his Metaphysics, Aristotle pointed out that "we understand those things best that we see grow from their very beginnings," but I disagree. After all, he provides little empirical evidence for this claim; when he does, it does not replicate—adding yet another example to the crisis in psychological research. You may be asking yourself, so why start with Aristotle? Establishing yourself as a bigger man is always a good idea by taking down one of the biggest. We learn this in the classic Spaghetti Westerns by Sergio Leone, where Clint Eastwood filled the screen but said nothing and added further evidence that smoking a cigar and wearing leather chaps is cool, way cool. If you are unsure what a Spaghetti Western is or why it is vital to my talk, ask an old Italian like Steve Sireci. Through the lens of the construct of egomania rageosis (Woolsey, 1987), my presentation aims to interrogate myself about what I consider our discipline's big questions, perhaps the biggest question of all: when I think, what do I think about measurement validity? I will address this question in a series of disconnected reflective narratives concerning my many selves, the boundaries of which are boundless, and my intimacy with those closest and farthest from me. To paraphrase Woolsey, the persons described in this presentation are entirely imaginary and bear no resemblance to any of my colleagues, ex-lovers, current girlfriends, past, present, or future spouses, living or dead. Threats of litigation should be addressed to the second author." (p. 3380) Based on the theory of learning that insists we break the subject down to rebuild them in one's image (Zumbo & Sycophant, 2005; Zumbo & Rigatoni, 2027; Zumbo, Vino, & Vongole, 1999), my presentation will be a success in a Latourian sense if you leave more confused about validity, overwhelmed by the possibilities, and have a nagging sense that we have spent over a century making impenetrable, the obvious, and straightforward (Zumbo, 2023).

WEDNESDAY 3 JULY
Session 8.3 SYMPOSIUM
Topic: Validity theory in testing, psychological assessment and survey research

660. Is there a natural law of validity or why Marlon Brando was so damn good as Don Corleone in The Godfather? (Validity theory in testing, psychological assessment and survey research)

Hudson Golino
University of Virginia

The current presentation will discuss if there is a natural law on validity. This delicious intellectual dish requires three cups of quantum information theory and a tablespoon of network science. To add even more flavor to the presentation, I will briefly show why John von Neumann and Satosi Watanabe developed information theory before Claude Shannon, maybe generating a diplomatic conflict while attempting to do so. The attendees will probably be shocked by my talk, which covers quantum information theory and validity simultaneously (especially if any physicists are attending the conference). Anticipating any possible conflict, I will show how Von Neumann's Entropy can be estimated in psychological data (and someone might have a heart attack in the conference room, so please have some physician on call). Then, I will try to answer the first question of the title: Is there a natural law of validity (spoiler, yes), and will work hard to answer the second question (why Marlon Brandon was so damn good as Don Corleone in The Godfather?) while connecting the later with the former. The current presentation aims to make you an offer you can't refuse. (Disclaimer: no large language model was used in writing this abstract, and my non-native English phrasal structure is the proof you need).

WEDNESDAY 3 JULY
Session 8.4
Topic: International assessment

288. TIMSS 2019 Equivalence Study: A Mixed-method Approach to Explore Assessment Mode Effects on Mathematics Performance in England

Liyuan Liu, Grace Grima, Sebastian Nastuta, Kevin Mason, Mish Mohan, Sarah Turner
Pearson Education UK

Introduction: TIMSS (Trends in International Mathematics and Science Study) started introducing a computerized version of the assessment at scale in 2019. In England, 9595 pupils from 368 schools participated in TIMSS 2019, with 6761 taking eTIMSS and 2834 taking paper-based tests. Although 80% of items were the same in both modes in TIMSS 2019, Fishbein et al. (2018), in their pilot study, identified heightened challenge levels in on-screen items, attributing this contrast to various factors like presentation style, typing proficiency, content complexity, and question types. Methodology: This study employs a two-phase mixed methodology to delve into 9595 Year 9 pupils' mathematical performance across assessment modes in England. Phase 1 involves analysing TIMSS 2019 student achievement data through t-tests, correct response ratios, and Item Response Theory (IRT) analyses. It aims to uncover overall mode effects and assess individual question difficulty levels and differences across on-screen and paper-based tests, focusing on four content domains. Phase 2 conducts focused interviews with TIMSS data scorers and mathematics experts, aiming to grasp pupils' interactions with the dual assessment modes deeply. Results &Implications: 1.While on-screen assessments were marginally tougher, statistically, the variance wasn't significant. 2. Empirical results highlighted substantial achievement gaps in items needing annotation, graphs, and mathematical work. Interviews underscored that question construct and cognitive demands might alter between modes despite apparent format similarities. 3. Familiarizing with digital tools (in-built calculators, etc.) may boost pupils' confidence during assessments. This research unpacks the practical considerations of dual assessment modes, suggesting strategies to mitigate their impact. It stresses the need for further exploration into these modes.

WEDNESDAY 3 JULY
Session 8.4
Topic: International assessment

422. Estimating Missing Home Socioeconomic Status in PIRLS using Student and School Questionnaire Data: A MICE Simulation.

João Marôco
William James Centre for Research, ISPA - Instituto Universitário. Portugal
Matthias von Davier
TIMSS and PIRLS International Research Center, Boston College. USA

Framework International large-scale student assessments (e.g., TIMSS, PIRLS, PISA) intentionally have missing data for estimating student achievement. While design-induced missing responses are less problematic, non-response due to students or parents omitting answers poses challenges for predictions. Objectives Explore the feasibility of imputing Home Socioeconomic Status (HSES) in PIRLS 2021 using Multivariate Imputation by Chained Equations (MICE) based on students' and schools' data. Sample Data includes PIRLS 2021 students' and schools' questionnaires, along with reading Rasch scores estimates. Methodology Utilized various MICE algorithms (e.g., pmm, midastouch, cart, rf, norm, lasso, stdwgt pmm) for imputations. Simulated different missingness levels (10-60% MCAR) on a 5,500-student sample. Country and region imputations were also performed using pmm and weighted pmm. Results Replicating imputed HSES data became less reliable with increasing MCAR data (10-60%), dropping from 36% explained variance to less than 16%. pmm showed superiority in accuracy and time efficiency. MICE imputations exhibited shrinkage, overestimating HSES for values below M-1SD and underestimating for values above M+1SD. Imputation efficiency varied by country, emphasizing the importance of data distribution and questionnaire completeness. Weighted pmm enhanced reliability within countries, especially for asymmetrical distributions and non-random missing data. Implications Researchers using MICE for HSES imputations should consider appropriate student weights, mindful of potential overestimation or underestimation, especially for values far from the mean. pmm is effective for symmetrical, Gaussian-like distributions, particularly with low to moderately missing data (up to 40% MCAR) clustered around the mean. Weighted pmm imputations using students' total weights significantly improve imputation quality, particularly for countries with asymmetrical distributions and non-random missing data

WEDNESDAY 3 JULY
Session 8.4
Topic: International assessment

820. International Assessment Reform-A Case Study in Egypt

Sharon Hague
Pearson Assessments
Jon Twing
University of Sydney

As one can imagine, transforming national examinations from paper and pen to an online mode of delivery represents quite a challenge. This challenge is even greater in the international contexts when the driver of change is education reform—where issues like culture, public policy, infrastructure, and curriculum all impact or are part of the transformation. Recently, part of Egypt's 2030 vision development strategy around social transformation at national level, focuses on its education reform efforts which they call vital. As part of Egypt's "Education 2.0" vision, Egypt secured financing from The World Bank to implement this massive reform effort. One of the goals of this reform effort is to improve educational learning outcomes for students and improve student engagement with education. Pearson is partnering with Egypt and The World Bank in implementing this reform effort and this paper will explain, in a case study approach, the challenges of such an effort. Other international jurisdictions should benefit from the lessons learned in Egypt as they consider their own reform efforts, goals, and objectives. Some of the implementation challenges encountered and addressed, as discussed in the paper, included: Rigor—adding real rigor to the assessments when historically, the majority of pupils sitting the exam passed (e.g., 40 percent of the candidates were earning 90 percent or more of the marks); Timing—rapid policy change required rapid implementation with all jurisdictions taking all subject matter examinations resulting in 1.5 million test takers the first year; Constructs—In the past, most teaching was rote memorization of skills and knowledge and the new exams moved to application; Tutoring—In the past a large and very lucrative tutoring business thrived on test preparation and the changes would disrupt this, causing political, policy and public push back on the reform efforts; Technology—when the project started, Egypt ranked 170th out of 200 regarding internet connectivity and bandwidth. Yet, both assessment and instruction were to make use of cloud-based delivery resources for all parts of the country. More specifics regarding how these challenges were overcome as well as other and still ongoing reform issues will be discussed as the project continues. For example, in 2021 40 different examinations were implemented in high schools to over 1.6 million students, this included "retakes" as well as "home-based" examinations, all of which needed the infrastructure and connectivity to meet reporting timelines. Plans to maintain and improve this reform effort going forward will also be discussed.

WEDNESDAY 3 JULY
Session 8.5
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

571. Validation of AQoL-8D measures: a health-related quality of life questionnaire for adult patients referred for otolaryngology

Anita Obrycka
1. World Hearing Center, Institute of Physiology and Pathology of Hearing, Kajetany/ Warsaw, Poland
Jose-Luis Padilla
2. Dept. of Methodology of Behavioral Sciences, University of Granada, Spain
Artur Lorens, Piotr Skarzynski, Henryk Skarzynski
World Hearing Center

Theoretical framework Many otolaryngology problems e.g. hearing loss, vertigo, tinnitus are often associated with long-lasting disabilities which usually have severe psycho-social consequences. Revision of generic health related quality of life HRQoL questionnaires shows that AQoL-8D is the instrument that has items suitable to capture "quality-of-life" domains important for otolaryngology. Objectives The study objective was to obtain validity evidence of AQoL-8D measures in adult population of patients referred to otolaryngology clinic. Sample 463 Polish patients age18–80 years with otolaryngological conditions. Methodology Patients were assessed with the AQoL-8D, SF-6D, and SWLS questionnaires. We investigated the item content-relevance, factor structure by means of Confirmatory Factor Analysis, corrected item–total correlations, Cronbach's alpha, Pearson correlation of the AQoL-8D scores with results from SF-6D and from the SWLS questionnaires. Finally, ANOVA was used to test the AQoL-8D ability to group the HRQoL of patients in terms of their otolaryngological management type. Results The median score of item content-relevance was 5.0 for all AQoL-8D items. Confirmatory Factor Analysis revealed the following fit indices: CFI = 0.81; TLI = 0.80; and RMSEA = 0.07. Cronbach's alpha for AQoL-8D dimensions ranged from 0.48 to 0.79. Mean item–total correlations over all dimensions, super dimensions, and the instrument overall were higher than 0.3. There was a significant Pearson correlation between the results obtained with AQoL-8D and SF-6D r = 0.68, and with AQoL-8D and SWLS r = 0.43. A one-way ANOVA showed a significant effect of management type on HRQoL as measured by AQoL-8D $F_{4,458} = 6.12$, $p < 0.001$ Implications AQoL-8D provides valid and reliable measures of HRQoL in patients undergoing otolaryngological treatment. With AQoL-8D it is possible to make general comparisons of otolaryngology outcomes with those from other subspecialties.

WEDNESDAY 3 JULY
Session 8.5
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

584. Developing culturally sensitive and linguistically accurate versions of a music-based assessment for adults with disorders of consciousness

Wendy Magee
Temple University, Philadelphia, USA
Camila Acosta Gonçalves
Universidade Federal Do Rio De Janeiro, Brasil
Anna Menén Sánchez
Institut Guttmann, Barcelona, Spain
Jingwen Zhang
Suny New Pañtz, NY, USA

Disorders of Consciousness (DoC) following acquired brain injury result in complex cognitive, sensory, and motor impairments. Determining patient awareness is the priority to guide future care. Language-based assessments are often insensitive to the subtle behaviours that might determine awareness. Music is emerging as a potential assessment and treatment modality to enhance responsiveness and identify awareness in DoC. The Music Therapy Assessment Tool for Awareness in Disorders of Consciousness (MATADOC) is a standardized music-based diagnostic assessment for use with DoC populations to guide treatment. The English language version has good inter-rater and test-retest reliability (Magee et al., 2014) and significant moderate agreement with the criterion standard (Magee et al., 2023). The protocol uses patient-preferred live music in five procedures to elicit behavioral responsiveness that is rated across 14 items following manualized guidelines. The MATADOC is used by clinicians trained in its use in 32 countries. We present work to develop culturally sensitive and linguistically accurate versions of the MATADOC protocol and documentation translated into Mandarin, Spanish and Brazilian Portuguese. Following the ITC (2017) guidelines, expert panels for each project refined forward translations and sensitized the live music protocol to cultural practices including recommendations for the inclusion of musical instruments and genres pertaining to each culture. Following a back translation, a small try-out trial in China demonstrated good clinical utility with promising inter-rater reliability and opportunities for family engagement. While constructs pertinent to DoC are consistent across cultures, linguistic consensus on essential terms in DoC assessment ("awareness", "arousal") within each culture proved challenging. Further testing of the revised protocols to optimize cultural relevance within collectivist cultures will enhance applicability of the translations

WEDNESDAY 3 JULY
Session 8.5
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

637. What is psychometric norm? Contemporary challenges for cultural adaptation of questionnaires for the assessment of mental disorders in adolescents

Joanna Stanczak, Iwona Bac, Emilia Wroclawska-Warchala, Radoslaw Wujcik
Pracownia Testow Psychologicznych PTP

The first half of the twenties of the 21st century is a time when difficult, unexpected, and still not fully understood events dominate both locally and globally. The pandemic that has engulfed the entire world, erupting armed conflicts, societal polarization, the growing strength of extremely nationalist movements, and the persistent lack of tolerance for diversity are just some of the challenging experiences present in the lives of young people. The impact of the pandemic and lockdown on the functioning of young people has been enormous. Social relationships and the sense of security have been disrupted. This is reflected in the growing emotional, psychological, and cognitive problems. In such a situation, how can standardization research of tools used to assess psychological disorders can be conducted and how can we define the psychometric norm? Should the fact that almost 1/3 of the respondents from the general population affirmatively respond to items on a scale related to psychopathological symptoms be considered the norm? In this paper we will discuss the results of two analyses: first, the analysis of atypical pattern received in scale F of MMPI-A-RF in general population of adolescents and predictors of the frequency of atypical responses, such as the results in BYI-2 and SENA scales measuring clinical symptoms. Multivariate logistic regression will be employed in the analysis. Then, we will discuss the differences between pre- and post-pandemic results of clinical symptoms scales of SENA, obtained in general population of adolescents before and after the pandemic. Post-pandemic data were collected in standardization studies of MMPI-A-RF. The research was conducted on youth aged 14-18; the standardization sample was quota-based and representative of the Polish youth population (N = 912). SENA and BYI-2 were used in this research as validation tools. Pre-pandemic data, used in second analysis, were collected in standardization studies of SENA (N = 200).

WEDNESDAY 3 JULY
Session 8.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

416. Applications of Artificial Intelligence to Support Human and Automated Scoring of Constructed Responses

Edward Wolfe
Pearson

Artificial intelligence (AI) allows computers to automate and improve decisions that have traditionally been delegated to humans. This seminar presents applications of AI to support human and automated scoring of constructed responses by three testing organizations. Presentation 1 summarizes an effort to support human scorers with feedback about and scores of examinee responses that was created by generative AI (GPT4). When raters received these supports during the scoring process, scoring speed increased and, in some cases, scoring accuracy improved. GPT4 scores were moderately correlated with scores from both an automated scoring engine (ASE) and human raters. The presenter also proposes a method for evaluating output from generative AI systems including evidence beyond what is typically considered in traditional evaluations of ASEs. Presentation 2 proposes an approach to handling responses that an ASE has difficulty scoring. The presenter describes an approach to optimize a hybrid scoring system that uses the best of ASE and human scorers to identify the most suitable and fairest outcome for examinees. That system utilizes an approach to training and evaluating ASE score confidence measures to predict when a response would be better scored by a human. The presenter also discusses methods for detecting responses that may trick an ASE. Presentation 3 addresses the challenge of training ASEs when some score categories are observed infrequently, where those who train ASEs must balance scoring costs with the need for training data (human scores). Data augmentation, creating responses that have characteristics similar to existing training examples via generative AI, is a potential remedy for this issue. The presenter will present results of experimentation that demonstrate that augmenting training data sets can lead to performance of an ASE on unseen data that is equivalent to or better than what is achieved with original responses.

WEDNESDAY 3 JULY
Session 8.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

423. An Application of Data Augmentation to Automated Scoring Engine Training (Artificial Intelligence in testing, psychological assessment and survey research)

Edward Wolfe, Justin Barber
Pearson

Automated scoring engines (ASEs) apply artificial intelligence (AI) to the scoring of constructed response (CR) assessment items via a computer algorithm that predicts the score that a human rater would assign to a particular response. ASEs are "trained" to make this prediction by determining the relationship between features of the CR and human-assigned scores that exists in training data. A common challenge encountered by those who develop ASEs is the fact that very few training examples might exist for score categories that are seldom observed, and those who train ASEs must balance scoring costs with the need for those (human scores) training data. Data augmentation, a process of creating responses that have characteristics similar to existing training examples via generative AI, is a potential remedy for this issue. We summarize the results of experimentation that evaluates the impact that augmenting training data sets can have on ASE performance on unseen data. Specifically, we utilized several generative AI methods to create simulated responses to CR items and subsequently trained an ASE using those responses to supplement the training process. In our experiment, we compared the agreement between human and ASE scores as measured by quadratic weighted kappa, varying the type of generative AI used, the proportion of simulated response in the ASE training data, the ASE training approach. Our results indicate that ASE scores from engines trained with data augmentation supplementation may achieve levels of agreement that are as good or even better than those demonstrated by human raters.

WEDNESDAY 3 JULY
Session 8.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

506. Curation of Validity Evidence for Applications of GPT4 in CR Scoring Systems (Artificial Intelligence in testing, psychological assessment and survey research)

Jodi Casabianca-Marshall, Vladimir Zubenko, Naim Alper
Educational Testing Service

Natural language processing (NLP) solutions for automatically scoring constructed responses (CR) are well established and used broadly in standardized testing. Some new artificial intelligence (AI) solutions, specifically generative AI such as GPT4, are not based on the same principles of NLP and have little-to-no transparency. However, they offer capabilities that were previously unavailable without extensive effort by experts. Since the future of assessment will likely maximize AI, it is important to understand how generative AI can assist with CR scoring. The goal of this presentation is to use findings from a study exploring the utility of GPT4 in the context of CR scoring to highlight how validity arguments for this approach to AI scoring should be specially curated. The study explored GPT4 in two applications, to generate: (i) feedback on responses to be used by human raters during scoring, and (ii) scores. We used written responses from four tests that were already scored by operational raters. For each response, GPT4 created bulleted feedback and a score based on prompting that included the scoring rubric, instructions, and exemplars. In a counterbalanced design, we collected ratings under three conditions–no assistance, AI feedback, and both AI feedback score. We also compared the GPT4 scores to scores from the operational AI scoring engines (NLP-based scores). Results showed that AI feedback increased scoring rates and improved accuracy in some cases. Scores from GPT4 were moderately correlated with NLP-based scores but had similar correlations with human ratings. The traditional types of validity evidence we might collect for AI scoring must be expanded for generative AI applications. For example, we might perform a qualitative analysis to understand what aspects of a response lead to differences in scores. We might also collect data over time to assess the reliability of the AI. We discuss the expansion and make suggestions for practitioners.

WEDNESDAY 3 JULY
Session 8.6 SYMPOSIUM
Topic: Artificial Intelligence in testing, psychological assessment and survey research

496. Developing & Evaluating a Hybrid-Marking System to Combine AI & Human Scoring Models (Artificial Intelligence in testing, psychological assessment and survey research)

Mark Brenchley
Cambridge University Press & Assessment

Whilst AI-trained automated essay scoring (AES) models offer many advantages to learners and assessment organizations alike, they bring with them several challenges (Williamson, Xi, & Breyer, 2012). A key such challenge is the scoring accuracy of AES models, a particular consideration given the status of examiner scores as underlying reference scores, the fact that examiner scores themselves can be subject to variation (Bennett & Zhang, 2016), and that some scripts are likely to be harder for an AES model to mark than others (Yan, Rupp & Foltz, 2020). One overarching response to this challenge is the use of a hybrid-marking system (e.g. Xu et al, 2020), where both human examiners and AES models are available, and confidence measures are used to make real-time decisions as to whether a script is most appropriately awarded the AES score or the human counterpart. In this presentation, we describe one such system, developed for a high-stakes Writing exam, and according to which candidates can receive the AES score for one or both of their responses. We detail various approaches to training and evaluating such a system. In particular, we detail a number of AES confidence measures and their evaluations to-date, highlighting some key dimensions based on these evaluations, including: the importance of both multi- and single-marked evaluation data, the value of importance sampling, and the relative salience of particular metrics such as RMSE and exact agreement. We further discuss the interaction of these confidence measures with several distinct automarker models, each with a different underlying performance, showing how confidence measures are an important consideration for determining which of an available set of automarkers performs best within a hybrid system. Finally, we discuss the interaction of the confidence measures with supplementary methods for detecting responses that could potentially "trick" an AES system, whether inadvertently or intentionally.

WEDNESDAY 3 JULY
Session 8.7
Topic: Construct or concept equivalence

267. The South African Personality Inventory (SAPI) across Cultures

Velichko Fetvadjiev
University of Amsterdam

The South African Personality Inventory (SAPI) was developed to address the need for fair assessment of personality across cultural groups in South Africa. The project is an example of the combined emic–etic approach to cross-cultural assessment, where, firstly, indigenous perspectives from distinct ethnolinguistic groups informed the development of a common personality model, and, secondly, this model was directly compared to presumed universal models such as the Big Five and the HEXACO. The SAPI contains large positive and negative social-relational factors that are empirically distinct from other models, suggesting blind spots in universal models. This talk presents the further cross-cultural replication of the SAPI model in English-speaking countries (New Zealand, Ireland) as well as its Dutch adaptation (The Netherlands) (combined N over 1000). The incremental validity of the SAPI model above the Big Five/HEXACO is examined for a range of behaviour outcomes in the social-relational domain, using both self- and peer-reports. Findings in all three countries suggest that the SAPI is applicable beyond South Africa. The results are discussed with reference to the integration of indigenous and cross-cultural perspectives in the assessment of personality.

WEDNESDAY 3 JULY
Session 8.7
Topic: Construct or concept equivalence

645. Assessment of collective efficacy and social cohesion in the context of policing

Miguel Inzunza
Unit of Police Work/Umeå University/Sweden

The level of collective efficacy in a neighborhood has been considered an important factor for advancing crime prevention initiatives enabling informal social control. In the present research a measure of collective efficacy and related constructs was adapted to the Colombian context. The purpose of the study was to assess the applicability of the measure in terms of construct validity, and construct equivalence. Research questions investigate the applicability of the measure when considering background information such as gender or physical context. Data was based on three sets of cross-sectional data with respondents from complex areas in several Colombian mayor cities. Data was analyzed adopting latent variable modeling. Findings show that the measure was useful in applied research to evaluate police strategies, one such example is increased police presence. Discussion concerns the measurement and the interpretation of findings.

WEDNESDAY 3 JULY
Session 8.7
Topic: Construct or concept equivalence

661. Exploring Construct Equivalence of Items in Standardized Performance Testing: A Meta-Analytical Perspective on Response Formats

Sonja Breuer, Thomas Scherndl, Tuulia M. Ortner
Paris Lodron University of Salzburg

Standardized performance tests wield significant influence across academic and professional realms, impacting students, educators, job seekers, researchers, and policymakers worldwide. Diverse response formats have characterized testing practices over recent decades. Among these, the closed-ended (CE) format, commonly known as multiple-choice testing, prevails in performance assessment, contrasting with various types of open-ended (OE) response formats. Despite observations of relatively strong correlations between scores from OE and CE formats, concerns emerged regarding the construct equivalence of test items across these formats. Therefore, we aimed to meta-analytically integrate existing research findings on the construct equivalence of test scores derived from varying types of OE (e.g., essay, written short-answer, practical task) and CE (e.g., true-false, single-choice, multiple-choice) response formats. Analyzing a total of 392 effect sizes from 102 primary studies, our findings indicated that scores from different response formats were related the most when written short-answer items were used rather than other open-ended items, and when closed-ended items with lower probabilities to guess and use test-taking strategies (e.g., multiple-choice) were used rather than closed-ended items with higher chances to guess and use test-taking strategies (e.g., true-false). Limitations and the results' implications for practitioners and researchers engaged in the field of standardized performance testing will be discussed.

WEDNESDAY 3 JULY
Session 8.8
Topic: Translation of tests, psychological assessment instruments and survey questionnaire/ Testing equivalence by psychometrics methods

70. What Does It Take To Make It: The Dark Side Of The Performing Arts

Melissa McMullan
Edinburgh Napier University

The aim of this study was to fill a gap in present psychological research by investigating the relationship between the Dark Triad personality traits (Machiavellianism, Narcissism, Psychopathy) and involvement and achievement in the Performing Arts. These constructs have been explored separately through creativity, theory of mind and emotional intelligence among other research which provides reason for correlation. This research was a within-subjects correlational design investigating the relationship between involvement and achievement in music, dance and acting and the Dark Triad. A survey including an adapted version of the Creativity of Achievement Questionnaire (CAQ) and the Short Dark Triad (SD3) were counterbalanced and distributed by snowballing techniques and random sampling using Qualtrics (N= 121). People with varied contact with the performing arts including, no experience, amateur and professional were included. This study used three Spearman's correlations and six linear regressions, the results show significant positive correlations between involvement and achievement in the performing arts and the Dark Triad traits. Multiple Linear regressions revealed that Narcissism positively predicted involvement and achievement in all three performing arts and Machiavellianism predicted achievement in dance. This research provides a significant novel contribution to current personality and creativity studies in psychology, as the results allow for future prediction of the traits observed in people who aspire to participate in the performing arts and who may be successful given their level of narcissistic traits. Overall, this research shows significant relationships and predictions between the Dark Triad personality traits and the performing arts.

WEDNESDAY 3 JULY
Session 8.8
Topic: Translation of tests, psychological assessment instruments and survey questionnaire/ Testing equivalence by psychometrics methods

624. A New View on the Standard Error of Equated Test Scores

Wim J. van der Linden
University of Twente

The standard error of observed-score equating is generally presented as an error in test equating caused by random sampling of the examinees from an assumed population. One of the landmark examples is the standard error for the randomly-equivalent-groups design derived in Lord (1982). It is argued that the cause of random error in an equating is not examinee sampling but measurement error in the observed scores that are equated. Consequently, to obtain the standard error of an equated score, the only necessary step is an adjustment of the standard error of the observed score for the impact of the equating transformation, an adjustment shown to be a factor with a simple analytic expression. However, as the standard errors used in the current literature generally ignore the existence of measurement error, no matter the type of adjustment, the necessary consequence is a standard error of equated scores always equal to zero. On the other hand, if we allow for the presence of measurement error, the result is a standard error of each equated score equal to the standard error of the score on the form to which it is equated.
Key words: equated scores; equating transformation; observed-score equating; standard error of equating.

WEDNESDAY 3 JULY
Session 8.8
Topic: Translation of tests, psychological assessment instruments and survey questionnaire/ Testing equivalence by psychometrics methods

628. Can computer adaptive version of students' ratings of instruction provide valid results?

Semih Topuz, Kadriye Belgin Demirus, Esra Kinay Cicek, Giray Berberoglu
Baskent University

Universities worldwide include the evaluation of instructional processes in their quality indicators. Student ratings often influence promotion decisions, but their validity is a subject of ongoing debate. External factors, such as course load and past experiences with instructors, may interfere with results. To ensure validity, it is important to consider these factors. Additionally, students' reluctance to fill out questionnaires and inattention in answering them are other factors that affect their validity. The amount of time that students spend on repetitive questionnaires may compromise the quality of data. To address this issue, reducing the number of questions and using different sets of items for various courses may enhance validity. One solution could be computerized adaptive testing. The study aims to simulate computer adaptive testing of students' ratings of instruction based on their preconceived opinions about the course, including course difficulty and expected letter grades at the end of the semester. The hypothesis is that starting the adaptive test administration with students' overall opinions will result in more accurate and valid results. This approach will also provide different sets of test items for different courses, which may improve the accuracy and validity of the rating results. In general in the present study the effect of considering students' preconceived opinions about the course difficulty and expected letter grades will enhance the adaptiveness of the whole process will be studied. Simulation will be carried out with different strategies by considering students responses on perception of course difficulty and the expected letter grades as starting point in adaptive process. The results will be compared with the results obtained in the full test with 16 items and the results of the simulations using fixed starting point and error estimation.

WEDNESDAY 3 JULY
Session 8.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

319. Equitable Selection into Initial Teacher Education Programs: The role of innovation, transparency and feedback

Janet Clinton, Katina Tan, Therese Hopfenbeck, John Hattie, Laura Smith
University of Melbourne

Teachers shape the future, and initial teacher education plays a critical role in ensuring that our future teachers can practice effectively in ever-changing, complex and diverse contexts. To this end, selection into initial teacher education (ITE) programs continues to be an area of interest, debate and scrutiny. There is an international recommendation that entrants to ITE programs be selected through sophisticated and transparent approaches that consider both academic and non-academic characteristics of prospective teacher candidates. One such approach in the Australian context which has proven to be sufficiently robust and comprehensive is the Teacher Capability Assessment Tool (TCAT). TCAT is an online tool with standardised scales that embeds a range of factors including motivations for teaching, cognitive reasoning skills and non-cognitive domains such as disposition, self-regulation, communication style, fairness, and cultural sensitivity. Following the selection process, candidates are provided individual feedback for reflection and teacher educators are provided cohort results to ensure impact. This symposium presents three perspectives relating to entry assessment for ITE. Research findings from the TCAT database of over 18,000 teacher candidates The first presentation is a grounding of TCAT and interrogates teacher disposition assessment and implications for preparing all teachers. The second focuses on issues of online assessment security, the advent of technology and the importance of innovative measurement techniques to ensure fair and equitable access. The final explores the feedback framework adopted by TCAT and its implication for potential teacher candidates, ITE providers and its impact on policy. In addition, future directions to support the dynamic, high-stakes ITE assessments are discussed. It is argued that we need to consider the intricate balance of robust teacher selection assessment and the implications for equity and fairness.

WEDNESDAY 3 JULY
Session 8.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

813. Assessing the competencies and characteristics of prospective teachers: A fair selection process? (Validity and fairness in cross-cultural testing, psychological assessment and survey research)

Janet Clinton, Laura Smith
University of Melbourne

The Teacher Capability Assessment Tool (TCAT) is a web-based assessment of the competencies, characteristics and attributes of individuals applying for initial teacher education programs. The tool gathers comprehensive information about a candidate's cognitive ability, personal characteristics, disposition, social interaction, cultural sensitivity, and self-awareness. The dimensions measured in TCAT are associated with student outcomes and teacher performance. Measuring these domains, rather than relying on interviews or essays, eliminates the potential for stereotyping or bias in the selection process; thus breaking down barriers to entry. This paper presents findings from the analysis of a large dataset of prospective teachers extracted from TCAT to demonstrate how measuring these domains contributes to fairness in teacher selection and in the classroom. This selection tool provides an opportunity to assess how a prospective teacher intends to deal with the moral and ethical nature of teaching and their sensitivity to various cultural issues and diverse contexts. Further, results suggest a relationship between teacher candidates' wellbeing and health-related dispositions and their intended behaviour relating to cultural sensitivity and fairness, values and ethics. The challenges of administering such an assessment in the current environment are raised. It is argued that utilizing candidates' dispositions, capabilities, and predicted behavior as a tool for selection into ITE has implications for barriers to entering the profession, teacher retention, and culturally sensitive classrooms. As such, a rigorous and fair assessment approach is essential.

WEDNESDAY 3 JULY
Session 8.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

814. Technology and Contemporary Assessment: The challenges for Academic Integrity (Validity and fairness in cross-cultural testing, psychological assessment and survey research)

Therese Hopfenbeck, Katina Tan
University of Melbourne

Amidst significant innovation leaps in areas such as artificial intelligence, machine learning, gamification, digital social networks and 'big-data', the fundamental consideration of any assessment, regardless of modality, remains the same – is the assessment valid, reliable and fair? The Teacher Capability Assessment Tool (TCAT) is an assessment that is administered online. In this environment, it is essential to consider academic integrity and how various assessment frameworks can assist in preventing and detecting academic dishonesty. Online academic dishonesty can take the form of impersonation, forbidden aids, collusion, plagiarism and gaming the system which we argue are threats to TCAT's integrity and validity. While web-based assessment platforms are now commonplace, and much is known about the opportunities and pitfalls associated with technology-based assessments, new boundaries are continually pushed at a fast-evolving pace. We discuss how TCAT uses machine learning to scale and increase efficiency, link data across multiple platforms, create user dashboards, provide feedback and support, and significantly utilising contemporary assessment approaches to prevent and detect potential academic dishonesty.

WEDNESDAY 3 JULY
Session 8.9 SYMPOSIUM
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

815. Developing evaluative thinking through sound assessment processes (Validity and fairness in cross-cultural testing, psychological assessment and survey research)

Janet Clinton, John Hattie
University of Melbourne

Generally, models of sound assessment practices increasingly emphasise assessment's formative role. Along with the notion of learning progress, assessment now must support not only sound judgments about competence but also generate meaningful feedback to guide continuous learning. Reconciling the tension between the assessments' focus on judgment and decision-making and feedback focus on growth and development represents a key challenge for teacher educators in initial teacher education. The TCAT is specifically designed to ensure that it adds value to the professional journey of pre-service teachers. It utilises an evaluative thinking lens to ensure impact. This paper presents literature in relation to perspectives on evaluation, assessment and feedback in the frame of selection into initial teacher education programs. The principles and use of feedback are discussed, and it is suggested that providing an opportunity to engage in formative assessment contributes to pre-service teachers' ongoing professional growth and their self-reflective capacity as teachers. Further, this approach models the ideals of the continuity of assessment.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

786. Are we really all that different? Examining Country and Gender Differential Item Functioning of the PROMIS Anxiety-8a.

Amanda Dumoulin, Xuyan Tang, Sophie Ma Zhu, Anita M. Hubley
University of British Columbia

Framework: Mental health concerns during and after the COVID-19 pandemic have heightened the need for screening measures with strong psychometric properties. The Patient-Reported Outcomes Measurement Information System (PROMIS) is an increasingly popular set of person-centred measures that evaluates physical, mental, and social health. The PROMIS Anxiety-8a measures fear, anxious misery, hyperarousal, and somatic symptoms of anxiety using a 5-point Likert-type scale. Before using this U.S.-based measure with Canadian samples, it is important to examine whether different groups might respond differently to some items despite being matched on the latent variable. Objective: The purpose of this study was to examine the differential item functioning (DIF) of the PROMIS Anxiety-8a in community samples of (a) Canadian vs. American adults, and (b) men vs. women. Sample: Participants were 515 adults (53% women, 45% men, 2% other) ages 18 to 81 years (M=44.4, SD=15.4) residing in Canada (55%) or the US (45%). Methods: Participants completed several health and wellbeing measures in an online survey; our focus was on the Anxiety-8a. We used confirmatory factor analyses (CFA), taking the ordered categorical nature of the responses into account, to examine the assumption of unidimensionality in all groups. Model fit was evaluated using robust CFI & TLI >.90 and RMSEA & SRMR

WEDNESDAY 3 JULY
Poster Session 4
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

124. Hazard PredictionTest to assess drivers who suffered a stroke and healthy drivers.

Candida Castro, Daniel Salazar-Frías, Lucía Laffarga, Ana Szot
CIMCYC (Mind, Brain and Behaviour Research Centre), Faculty of Psychology, University of Granada, Spain
María Rodríguez Bailón
Universidad de Málaga

In the present study, we analysed differences in Hazard Prediction (which measures a driver's situational awareness and involves perception, comprehension and projection of a situation) between persons who suffered a stroke, differentiated by the affected hemisphere. The results showed that drivers with a left side stroke obtained poorer accuracy percentages in the Hazard Prediction test (Castro et al. 2021) than drivers with a right side stroke and healthy drivers. This effect was especially pronounced in unfamiliar localities (U.K. videos). One possible explanation of our finding (drivers with a left side stroke obtained poorest accuracy percentages) is that an injury in the left hemisphere affects one's ability to analyse and one's analytical thinking. In fact, various studies found that the left hemisphere appears to be more involved in executive processing and in the processing of local information, which can be important for recognizing objects, whereas the right hemisphere appears to become more active when more global characteristics are involved, such as the general navigation of one's surroundings (Chokron, et al., 2000). Additionally, other researchers revealed that the prediction of road scores following a left side stroke involves several cognitive factors, such as processing speed and executive dysfunction (Devos, et al. 2014). In fact, in contrast to Sasaki et al. (2019), the present study also found differences in response times in the Trail Making Test-B (TMT-B) between drivers who suffered brain damage in different hemispheres, with greater response latency in patients with damage to the left hemisphere compared to the response latency in patients with damage to the right hemisphere. This could indicate that the Hazard Prediction test is more sensitive than the traditional Hazard Perception test in discriminating between drivers with different types of brain damage.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

382. Psychometric Properties of the Spanish Version of the Cannabis Refusal Self-Efficacy Questionnaire (S-CRSEQ-13) among Young Adults who use cannabis.

Lucía Vélez-Pérez11/
Department of Clinical and Experimental Psychology, University of Huelva, Avda. Fuerzas Armadas, s/n, Huelva 21071, Spain
Bella María González Ponce, Nehemías Romero-Pérez
Department of Clinical and Experimental Psychology, University of Huelva (Spain)
Adrian J. Bravo
Department of Psychological Sciences, William & Mary, Williamsburg VA, United States
Fermín Fernández-Calderón
Department of Clinical and Experimental Psychology, University of Huelva (Spain)

Background: Cannabis refusal self-efficacy, defined as the ability of people who use cannabis to refuse this substance, is a central construct of social cognitive theories that has been shown to be useful in determining cannabis use and cannabis use disorder. The Cannabis Refusal Self-Efficacy Questionnaire (CRSEQ) has shown utility for evaluating this construct. However, a Spanish adaptation of this measure is not available. Objectives: We aimed to provide a Spanish version of the CRSEQ-13 and examine its psychometric properties. Specifically, we examined its reliability and sources of validity evidence (structural and concurrent validity) in a sample of young adults who use cannabis. Methods: Targeted sampling procedure was used to access a community sample of young adults (n=612; mean age=21.04, SD=2.16; 38.4% female). For the adaptation of the instrument, several experts were involved in a process of translation and back translation. The participants completed the Spanish CRSEQ at baseline and at three months follow-up (follow-up participants=505), along with measures of cannabis use, cannabis-related problems, and cannabis protective behavioral strategies (PBS). Results: Confirmatory factor analysis showed that the original structure of three factors (emotional relief, opportunistic, and social facilitation) was supported, both at baseline (CFI = .928, RMSEA = .078, SRMR = .056) and follow-up (CFI = .922, RMSEA = .086, SRMR = .063), after deleting item 4 due to its low factor loading. Moreover, internal consistency reliability ranged between .59-.94 and evidence of validity was provided according to the expected relationships with other variables. Conclusion/Implications: Our results support the utility of the CRSEQ among Spanish young adults who use cannabis. This work was supported by the Agencia Estatal de Investigación (Ministerio de Ciencia e Innovación, Spain) under Grant Number PID2020-118229RB-I00 (PI: Fermín Fernández Calderón).

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

601. Assessing Teachers' Rational Number Knowledge: An Evaluation of Initial Validity Evidence.

Joanne Joo, Leanne Ketterlin Geller
Southern Methodist University
Sarah Powell
The University of Texas at Austin
Erica Lembke
University of Missouri

Measuring teachers' mathematics content knowledge, particularly the knowledge pertaining to the topic of instruction, is crucial. As part of a larger project studying the effect of supplemental instruction focused on rational numbers to Grade 4 students, we measured teachers' knowledge of rational numbers. Ultimately, we aim to understand if teachers' knowledge affects student learning outcomes. While research in this area has shown modest effect sizes, the potential impact cannot be dismissed. The initial step in this process is the development of a technically adequate measure to assess teachers' rational number knowledge – Rational Number Assessment (RNA). This presentation reports on the psychometric analyses conducted to evaluate the RNA for use as a pre- and post-test measure of teachers' content knowledge. We strive to create a measure that demonstrates sensitivity to teachers' knowledge growth over time, thereby ensuring its effectiveness in evaluating teachers' competency in this important area. To examine the psychometric qualities of RNA, we collected data from 72 Grade 4 teachers based in Texas and Missouri. 93% of the teachers were female and 76% were White. They had an average of 12.88 years of teaching experience. We used a Rasch model to examine the item parameters and fit statistics. Item difficulty parameters ranged from -3.43 to 3.56; however, the mean person ability was approximately 0.80, indicating that the overall difficulty of the RNA was not well aligned with the sample. In future revisions, we aim to consider adding more difficult items to better measure the range of abilities. The point-measure correlations were within an acceptable range for all items (>.30). Five items had outfit statistics greater than 1.5, and thus will be dropped from future revisions to the RNA. All items had appropriate infit values. Cronbach alpha was .91. Implications for the validity of the uses and interpretations of the scores will be discussed.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

486. Psychometric properties of the Smartphone Addiction Scale (SAS-SV) in adolescents in Peru.

Joel Figueroa-Quiñones
Universidad Señor de Sipàn

Background: During 2022, it was estimated that 90% of the world's population would have at least one smartphone. This rise has far surpassed their original function as mere mobile phones, influencing the way we communicate, access information and manage our daily lives. Scientific literature has begun to explore the effects of smartphones on cognitive and emotional functioning. Objective: to evaluate the structural validity, reliability, and measurement invariance of SAS-SV in Peruvian adolescents. Materials and Methods: 1,274 adolescents of both sexes, between 12 and 17 years old, residing in the cities of the coast (Chimbote), mountains (Cajamarca), and jungle (Tarapoto) of Peru were evaluated. A confirmatory factor analysis was performed, and the measurement invariance according to age, region, and sex was evaluated using MIMIC (Multiple Indicator, Multiple Cause) models. Reliability was estimated through the Omega coefficient. Results: A 10-item univariate model was obtained with optimal goodness-of-fit indices (CFI= 0.98; TLI= 0.97; SRMR=0.06; RMSEA= 0.11). MIMIC models reported invariance for age, sex, and region groups (ΔCFI < .01, ΔRMSEA < .015). Reliability was optimal (Ω = .86). Conclusion: The Peruvian version of the SAS-SV (10 items) has shown adequate psychometric properties for use in the adolescent population.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

255. Analyzing Age-Based Measurement Invariance: A Study of the Basque Adaptation of the GPIUS-2 in Problematic Internet Use.

Jone Aliri, Olatz Goñi-Balentziaga, Nekane Balluerka, Arantxa Gorostiaga
University of the Basque Country UPV/EHU

Problematic internet use, defined as a behavior encompassing excessive, disproportionate, or inappropriate use of the Internet leading to distress, significant time consumption, and impairment of normal functioning across various crucial life domains, is emerging as a major issue in many developed countries (Kuss et al., 2013, 2014). The growing interest in exploring this phenomenon has led to the proliferation of assessment tools designed to evaluate this construct. One of the most promising questionnaires is the Generalized Problematic Internet Use Scale – 2 (GPIUS-2, Caplan 2010), specifically designed to assess the cognitive and behavioral aspects of problematic Internet use and its associated consequences. The present study aims to analyze the age-based measurement invariance of the Basque adaptation of the GPIUS-2. . The sample comprises adults (over 18 years) and adolescents (11 to 16 years) from the Basque Country (53.0% women, 46.7% men and 0.2% non-binary individuals). While the original GPIUS-2 features five first-order factors and a second-order factor, some adaptations, including the Basque version, propose a four-factor structure (Preference for social online interaction, Mood regulation, Negative outcomes, and Deficient self-Regulation). Once the instrument has been adapted to Basque, it is of great interest to analyze the invariance depending on the age of said instrument. For that purpose, multi-group confirmatory factor analysis was used to assess the measurement invariance of the four dimensions of the GPIUS-2 across two age groups. Results indicate a well-fitting constrained model (CFI=0.988; TLI=0.990; RMSEA=0.071), demonstrating successful age-based measurement invariance. These findings are crucial, as understanding problematic internet use across ages enhances the development of precise assessment tools, improves research quality, and facilitates personalized clinical interventions based on age.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

720. Competitive Latent Structures for the Comic Style Markers: Developing a Psychometrically Sound Short Version Using Spanish and US American Samples.

Jorge Torres-Marín, Ginés Navarro-Carrillo, Mariela Bustos-Ortega
University of Granada
Sonja Heintz,
University of Plymouth
Hugo Carretero-Dios
University of Granada

The Comic Style Markers (CSM) is a questionnaire that allows a fine-grained description of how people differ in the way they display humor in their daily lives. It includes 48 statements capturing eight interrelated, yet distinct comic styles: fun, irony, wit, sarcasm, benevolent humor, satire, nonsense humor, and cynicism. Despite the independent conceptual roots of these humorous domains, the analysis of the CSM scales' latent structure shows that their empirical distinction needs to be improved. Using the information derived from a competitive latent approach, including confirmatory factor analysis, bifactor analysis, and exploratory structural equation modeling, we proposed and validated a shorter 24-item version of the CSM in a large sample of 925 Spanish individuals (CSM-24-SP). This scale-refinement improved the psychometric differentiation of the eight comic styles without undermining the good internal consistency and the temporal stability of the CSM scores. Strong invariance was held for gender and age groups, and partial scalar invariance for countries also emerged using a sample of 318 U.S. American adults. Latent mean comparisons indicated that men were more likely to describe themselves as high in all comic styles than women, particularly in the use of sarcasm, cynicism, and satire. In terms of age-related differences, we only observed that fun scores were more pronounced among younger people. Moreover, whereas Spanish people showed greater levels of fun, irony, nonsense humor, and benevolent humor, U.S. Americans were more inclined to cynicism and, to a lesser extent, satire. Finally, regarding validity evidence based on relations to other variables, structural equation modeling corroborated a coherent nomological network for the CSM-24-SP, with dispositional expressions of benevolent humor (positively) and cynicism (negatively) outperforming other comic styles in accounting for individuals' well-being.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

142. Video Game Dependency Scale: A Reliability Generalization Meta-analysis.

Júlia Gisbert-Pérez, Elena Cejalvo, Manuel Martí-Vilar, Laura Badenes-Ribera
Universitat de València

The Video Game Dependence Scale (abbreviated as VGDS or CSAS in the German version) is used to assess Internet gaming disorder, both online and offline. This instrument comprises the nine DSM-5 criteria for the diagnosis of Internet gaming disorder and reflects a first estimate of the risk of developing the disease. The purpose of this study was to conduct a reliability generalization meta-analysis study to estimate the average reliability of CSAS scores and the degree of heterogeneity in reliability coefficients across different samples and contexts. A Reliability Generalization Meta-analysis study was carried out based on the REGEMA guidelines (Sánchez-Meca et al., 2021). The electronic search was performed in four databases, Web of Science (WoS), Scopus, Pubmed, and PsycInfo, as well as in Google Scholar. The search identified 19 studies that applied the VGDS scale. From the 19 articles identified, 14 articles reported some or several reliability estimates based on study-specific samples. A total of 16 Cronbach's alpha coefficients were provided, which were transformed by applying the formula proposed by Bonett (2002) to normalize their distributions and stabilize their variances. After, the average reliability coefficients and their confidence limits were back-transformed into the Cronbach's alpha coefficient metric. In the statistical analysis, a random effects model was applied to estimate the mean reliability coefficient and its 95% confidence interval with the improved method proposed by Hartung and Knapp (2001), and forest plot was constructed. The mean reliability of the VGDS total scores assessed as internal consistency using Cronbach's alpha coefficient was 0.925 (95% CI [.901,.942]). There was significant heterogeneity among Cronbach's alpha coefficients. According to psychometric theory, the VGDS is a reliable instrument to be used for exploratory purposes in the field and in clinical practice.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

498. Operational definitions and measurement of Externalizing: an integrative review and a new proposal.

Lidia Torres Rosado, Cinta Mancheño Velasco, Alberto Parrado González, Óscar Lozano Rojas
Universidad de Huelva

In the assessment of Externalizing psychopathology, various tests employing different operational definitions are utilized. Typically, these assessments are based on the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases (ICD). However, these taxonomies have continuously updated diagnostic criteria, leading to changes in the operational definitions of disorders falling under externalizing over time. In recent years, models such as the Alternative Model of Personality Disorders (AMPD-APA, 2013) and the Hierarchical Taxonomy of Psychopathology (HiTOP -Kotov et al., 2017) have emerged. Currently, there is a diversity of models underpinning the operational definition of externalizing disorders, raising questions about the comparability of measures derived from them. This work aims to provide a specialized literature review focused on operational definitions, dimensions and measurement instruments derived for externalizing psychopathology. Alongside this review, an integrative proposal is presented for measuring externalizing psychopathology, compatible with the diagnostic criteria of the DSM-5 classification system used in clinical practice and the HiTOP model, which is more commonly used in research. Regarding the analysis of operational definitions, ICD and DSM versions have provided more detailed descriptions of diagnostic criteria, facilitating the development of measurement instruments. However, the measurement of disorders from DSM/ICD is questioned. Recent models, such as HiTOP, attempt to overcome those criticisms. However, they also pose challenges for their measurement. A review of instruments under each approach shows incomplete coverage of Externalizing. Efforts to integrate ICD/DSM with other theoretical models of psychopathology and personality are still necessary. The integrative operational definition of Externalizing and its dimensions and traits provided may help in this regard.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

271. Voluntary Simplicity and Social Psychology: creation of a scale.

Luis Mundi López
Department of Social Psychology, Granada University, Spain
Chiara Ambrosio
Department of Psychology, University of Campania "Luigi Vanvitelli", Italy
Eva Moreno-Bella
Department of Social and Organizational Psychology, National University of Distance Education (UNED), Spain
Josefa Ruiz Romero,
Department of Social Psychology, Granada University, Spain
Andrea Velandia Morales
Department of Developmental and Educational Psychology, Granada University, Spain
Guillermo Willis Sánchez
Department of Social Psychology, Granada University, Spain

Voluntary Simplicity (VS) is defined as a set of practices adopted by individuals who voluntarily reduce their consumption and material possessions. Despite its relevance, there is limited literature that has examined VS, and the existing literature has issues regarding its assessment. Specifically, we argue that it is important to develop a measurement scale that differentiate between the different reasons behind these practices. After reviewing the literature, we suggest that there are three different reasons for adopting voluntary simplicity: environmental motivations focused on the desire to contribute to environmental protection and combat climate change by reducing the use of natural resources and waste generation; social motivations centered on the desire to contribute to achieving a more egalitarian and just society that respects labor rights and human rights; and personal motivations devoted to enhancing life satisfaction and well-being (Alexander and Ussher, 2012;Willis et al., 2023). Hence, this research aims to create a scale that properly assess VS while addressing the limitations of existing scales. This will enable the distinction of VS from other constructs, such as frugalism. Building on this foundation, our intention was to create a scale to measure this construct following the criteria specified by AERA, APA, and NCME. Initially, an expert judgment was conducted with six experts from different fields (e.g., inequality and consumption). Following this, and after verifying the functionality of the items, a study (N = 200) was conducted where both an EFA and CFA were performed. Results consistently showed a bifactorial —instead of a trifactorial— scale with personal and social motives loading in the first factor and environmental concerns loading in the second factor. We will discuss the importance of these results for the voluntary simplicity literature.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

295. Validation of the Spanish version of the Emotional Style Questionnaire.

Maria Dolores Lopez-Martinez, Maria Dolores Hidalgo-Montesinos
University of Murcia

Emotions are central and very important to the human experience. Thus, our usual emotional patterns have an impact on virtually all areas of our lives and are closely related intertwined with our well-being. The Emotional Style Questionnaire (ESQ) is based on a theoretical framework proposed by Davidson and Begley (2012), drawn from neuroscientific studies of emotion. Therefore, having instruments adapted into Spanish, that are easy to administer and that can be useful in assessing emotional patterns might be helpful as providing descriptions of specific undesirable emotional patterns. Thus, the aim of this study was to adapt and translate into Spanish the ESQ. The ESQ features 24 items self-report intended to capture how people vary across six dimensions that make up a healthy emotional life (Outlook, Resilience, Social Intuition, Self-Awareness, Sensitivity to Context and Attention). A sample of 249 participants between 18 and 63 years old (M = 23, SD = 7.7) completed the Spanish version of the Emotional Style Questionnaire (ESQ; Kesebir et al., 2020), General Health Questionnaire (GHQ-12; Goldberg & Williams, 1988) and the Big Five Inventory (BFI; John, et al., 1991). Socio-demographic data were also collected, being most participants female (61.8%), wage earner (53.3%), with secondary education (72.7%), and living with their own family (64.7%). Internal consistency indices of the six subscales were good. Regarding the evidence of internal validity, the analysis of dimensionality indicated a structure of six dimensions. As expected, the EQS scores were strongly correlated with general health measures. The Spanish version of the ESQ showed adequate psychometric properties being a suitable tool to assess emotional style within the Spanish context. Initial validity evidence is promising for the use of the adapted measure in further research and clinical practice. Finally, the Spanish version are facilitate the study of emotional styles at a cross-cultural level.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

794. Cross-Cultural Adaptation of the Perth Alexithymia Questionnaire (PAQ) for the Brazilian Context.

Maria Julia De Melo Amorim Venâncio, Cristiane Faiad de Moura
Universidade de Brasília

Alexithymia is characterized by difficulties in identifying and describing feelings, coupled with externally oriented thinking, constituting a transdiagnostic risk factor for psychopathologies. Additionally, alexithymia is present in approximately 50% of the autistic population. The Perth Alexithymia Questionnaire (PAQ) is a self-report measure consisting of 24 items, originally developed in English, aiming to assess alexithymia based on the attention-appraisal model (Preece et al., 2018). This study aimed to perform the cross-cultural adaptation of the PAQ for Brazilian Portuguese and provide evidence of content validity through evaluation by expert judges and individuals from the target population. The method of this study was structured into five stages: 1. obtaining permission for adaptation; 2. translation of the instrument; 3. synthesis and evaluation by experts; 4. back-translation and evaluation by experts; 5. evaluation by respondents from the target population. Universal Test Design principles were observed during the cross-cultural adaptation process, particularly considering the characteristics of potential autistic respondents (Cassidy et al., 2018; Williams et al., 2021). The questionnaire showed good evidence of content validity through judge's analysis (CVC = 0.90). Brazilian autistic respondents consulted did not report difficulties related to understanding the items. Further research is needed to provide validity evidence for the tool in the Brazilian population, as well as for specific clinical groups. It is hoped to contribute to the provision of an alexithymia tool grounded in a theoretical framework consistent with the most current literature (Preece et al., 2023), appropriate and adapted for the Brazilian population, considering Universal Test Design principles. The adapted version of PAQ has the potential to benefit research and psychological assessment processes, access to diagnoses, and clinical decision-making by clinical psychologists.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

369. Optimization of a self-regulation for learning online scale using IRT information.

Mariel Fernanda Musso
CIIPME (CONICET)- UADE (Argentina)
Eduardo C. Cascallar
KU Leuven (Belgium)

Self-regulated learning, which includes motivational beliefs and learning strategies, is crucial for academic success in online and blended learning environments. Some validated self-report measures have emerged to evaluate SRL in this context during the last decade. In particular, the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich et al.,1991, 1993) was often modified for online and blended learning contexts. However, it would be useful to have an analysis of the items' psychometric properties using IRT models. A major advantage of using IRT is the possibility of identifying items which maximize the information on the construct according to the purpose of the evaluation. This allows the possible insertion of specific items before, during, and after online instruction that could provide information for tutoring systems in real time. This study aims to analyze the psychometric properties of the MSLQ online version. The sample included 625 university students (female=60%), mean age= 29.3 (SD=10.85), enrolled in online courses of the Humanities track, during the academic year 2020-2021. A validated Spanish version was administered, adapted for use in online instruction, to measure Learning Strategies/ techniques (α=.95), Self-Efficacy/goals (α=.94), and Affect/Emotions (α=.87). Results of the IRT analysis show adequate discrimination parameters for most of the items in each dimension (>.60): Learning strategies (n= 26 items; a= 1.05; b= -.55), Self-efficacy/Goals (n= 16 items; a= 1.07; b= -.86), and Affect/Emotions (n= 12 items; a= .89; b= .17). The subset of items in the Learning Strategies and Self-efficacy dimensions showed good reliability of at least .90 for students with low to moderate values of ability. The subset of items in the Affect/Emotion dimension showed an acceptable reliability of at least .80 for students with moderate to high levels of this attribute. DIF analyses by gender and cognitive profiles will be reported.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

375. Psychometric properties of the EPPIP Scale in Health Sciences students from a Peruvian university.

Mercedes Merryl Jesus Peña, Antonella Alexandra Gallegos Arteaga, Rabbi Robinson Reyes Robles
Universidad Peruana Los Andes

Parental styles have been investigated for a long time, and a series of instruments have been designed to measure it. It is essential that the instruments be adapted to the geographic and sociocultural context of the populations. For this reason, it has become pertinent to establish the psychometric properties of the EPPIP scale in university students of the Faculty of Health Sciences-UPLA, 2019. The research was carried out with 984 students between the ages of 17 to 35 years old from the first to the ninth cycle of the Faculty of Health Sciences of the Universidad Peruana los Andes, using a non-probabilistic, stratified sampling. The results show that the EPPIP scale presents the reliability indexes are higher than 0.80 for the father and mother version. In each of the subscales, it presents reliability indices higher than 0.70. Regarding the discrimination index, it is also observed that the values of the items are higher than 0.30, which shows that the items have high representativeness for each scale and subscale. Subsequently, the confirmatory factor analysis showed a close adjustment to the optimum for the mother version: CFI = .891, GFI = .891, AGFI = .867, RMSEA = .063, RMR = .010, NNFI = .814 and TLI =. 814. And in the parent version: CFI = .856, GFI = .906, AGFI = .885, RMSEA = .058, RMR = .009, NNFI = .814 and TLI = .839 in the two-dimensional independent model. Finally, the scales of the scale were elaborated for an objective classification of the administered, concluding the EPPIP scale presents adequate psychometric properties.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

169. The Kuwaiti-Arabic Trait Emotional Intelligence Questionnaire-Short Form: The Adaptation and Validation of the TEIQue-SF in Kuwait.

Nasser Hasan
Kuwait University/Kuwait
Konstantinos Petrides
UCL/UK

This study aims to cross-culturally adapt and examine the psychometric properties of the Kuwaiti-Arabic Trait Emotional Intelligence – Short Form (TEIQue-SF) through Structural Equation Modelling (SEM). The adapted measure was administered to 1458 university students in Kuwait together with the Kuwaiti NEO-FFI. Reliability estimates for all TEIQue-SF variables were within the acceptable range, with the exception of certain factors as expected by the literature. SEM results suggested that the bi-factor ESEM model fit the data for the TEIQue-SF. Evidence for criterion validity was obtained through relationships between the TEIQue-SF with the Big Five Personality variables. The results suggested that the Kuwaiti-Arabic TEIQue-SF can be considered as a reliable and valid measure to study trait EI with the Kuwaiti population, and consequently, allow for cross-cultural trait EI comparisons.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

715. The Illness Management and Recovery Scale: Translation and validation study of the Spanish version.

Nuria Martín Ordiales, Mª Dolores Hidalgo Montesinos
University of Murcia
Maite Barrios Cerrejon
University of Barcelona
Mª Pilar Martín Chaparro
University of Murcia

The present study aimed to translate, adapt, and assess the psychometric properties of the Illness Management and Recovery (IMR) Scale's Clinician and Client versions into Spanish for use among Spanish mental health clinicians and service users. The translation process for both, the Clinician and Client versions of the IMR scale, adhered to the International Test Commission's 2018 guidelines, ensuring cultural and linguistic appropriateness for the Spanish-speaking audience. The study involved a sample of 49 mental health professionals who provided 166 responses to the IMR-Clinician scale, and 172 users of mental health services aged 19-68 (M = 47.24; DT = 9.43) with diagnosis of severe mental disorder (schizophrenia 61%, bipolar disorder 8.7%, schizoaffective disorder 9.3%, and others 6.8%) who completed the IMR-Client scale. Additionally, mental health service users also completed the Recovery Assessment Scale and the Dispositional Hope Scale. The adapted Spanish versions showed considerable adequacy, with a three-factor structure consistent across both versions, encompassing Management and Personal Goals; Coping, Recovery and Symptoms; and Effective Use of Medication and Substance Abuse factors. This structure aligns with the dimensional structure observed in previous studies. While the Coping, Recovery, and Symptoms factor demonstrated adequate internal consistency, the other two factors exhibited lower reliability. The correlations with other mental health recovery measures supported the scale's external validity. Despite some limitations, such as convenience sampling, the findings endorse the utility of the Spanish IMR Scale for assessing mental health recovery, presenting a unique alternative to measure the perspectives of clinicians and mental health service users. This research contributes significantly to the tools available for evaluating illness management and recovery among Spanish-speaking populations affected by severe mental disorders.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

494. The Screen for Cognitive Impairment in Psychiatry-Spanish in older adults: Preliminary analysis of its dimensionality and internal consistency.

Oscar Pino
Hospital Benito Menni, Barcelona, Spain
Georgina Guilera
University of Barcelona, Barcelona, Spain
Vanessa Sanz, Maria Guallart
University of Zaragoza, Zaragoza, Spain
Emilio Rojo
International University of Catalonia, Barcelona, Spain
Juana Gómez-Benito
University of Barcelona, Barcelona, Spain

The Screen for Cognitive Impairment in Psychiatry (SCIP) comprises five subtests for measuring verbal learning (immediate and delayed), working memory, verbal fluency, and psychomotor speed. The test has three alternative forms, each requiring less than 15 minutes to complete. The psychometric properties of the Spanish version (SCIP-S) have been extensively studied in young and middle-aged adults (18 – 54 years old) from the general population and in people diagnosed with severe mental disorders. However, its performance in older adults has not been analyzed to date. The aim of the study is to analyse the dimensional structure and internal consistency of the SCIP-S in a preliminary sample of older adults. Participants were individuals aged at least 55 years old living in private residential centres in the area of Zaragoza, Spain. Exclusion criteria included the presence of a relevant cognitive impairment or difficulties in understanding, as well as the presence of a severe or decompensated somatic or neurological disease. One of the three alternate forms of the SCIP-S was administered, along with the Spanish version of the Mini-Mental State Examination. The sample comprised 25 men (26.3%) and 70 women (73.7%), with a mean age of 74.13 (SD = 12.85, range 55 – 98). A total of 56.8% had completed primary school studies, while the remaining 43.2% had secondary school studies. Principal axis analysis revealed a one-factor structure accounting for 63.8% of the total variance. Factor loadings ranged from .653 (VLT-D) to .904 (VLT-I). Internal consistency, as measured by Cronbach's alpha, reached a value of .87. The total SCIP score was strongly related to the MMSE score, with a correlation coefficient of .70. Based on these findings, the SCIP-S demonstrates a one-factor structure and high internal consistency in the older adult population, supporting its potential utility in clinical and research settings focused on cognitive health in older adults.

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

360. Psychometric properties of ITQ and IPQ: Adaptations to the Spanish context to measure efficacy of Virtual Environments to generate emotional states.

Pablo Doncel, Miguel-Angel Muñoz, Maria Blasa Sanchez-Barrera, Pedro Garcia-Fernandez, Francisco Gómez
Universidad de Granada
Jolanda Tromp
Oswego. State University of New York
Daniel Salazar-Frías, Andreea Ionela Dinu, Candida Castro
Universidad de Granada

Virtual Reality (VR) is a promising tool for Psychology. Before applying VR. Instruments are needed to assess: 1. The general propensity of people to immerse themselves in VR; 2. The capacity of the virtual environment to produce in people a subjective experience of being in a place or environment different from the one in which they are physically. We adapted ITQ (Immersive Tendencies Questionnaire; Witmer & Singer, 1998) and IPQ (Igroup Presence Questionnaire; Schubert et al., 2001) to the Spanish language and context. 288 participants ($\bar{x}$ Age=23.28; 133 men) completed the questionnaires. ITQ bifactor structure (Focus and Involvement; 10 items; Rozsá, 2022) was replicated, obtaining acceptable goodness of fit indices by CFA (2[34]=69.79, p

WEDNESDAY 3 JULY
Poster Session 4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

782. Perceived utility of Q-matrices in the translation of diagnostic assessments: Translating an academic literacy test used for diagnostic purposes.

Sanet Steyn
University of Cape Town

The effectiveness of a diagnostic evaluation in educational environments is not solely dependent on its precision but also on its capacity to offer detailed insights into a student's performance within a specific area. Cognitive diagnostic modelling (CDM), particularly the utilization of Q-matrices to delineate attributes associated with individual test items, has been extensively studied to validate diagnostic tools. Despite its reliance on human judgment, the framework it provides for capturing test item specifications remains valuable. In high-stakes assessment scenarios, where the comparability of test forms is highly prized, ensuring parity between forms utilizing different languages as the assessment medium requires thoughtful consideration of various frameworks that can facilitate assessment development, including the translation of parts of or full tests. The National Benchmark Tests (NBTs) are widely recognized high-stakes assessments in the South African higher education landscape and are administered in two languages, English and Afrikaans. These assessments, which cover academic literacy (AL), quantitative literacy (QL), and mathematics, are employed to evaluate students' preparedness for the academic challenges of higher education studies. The investigation presented in this poster examines feedback data from of a group of test developers and test translators involved in the development of the NBT AL and NBT QL tests focused on their perceptions of: 1) the accuracy of Q-matrices developed to complement existing test specifications; 2) the challenges they had in creating and using these item-attribute descriptions; and 3) the utility of these Q-matrices in facilitation the development, review and/or translation process.

FRIDAY 5 JULY
Session 9.1 SYMPOSIUM
Topic: Innovations in test development

529. MatriKS: a New Computerized Raven-like Test for the Efficient Assessment of Fluid Intelligence

Debora de Chiusole
University of Padua
Alice Bacherini
University of Perugia

The symposium focuses on introducing advanced methods and procedures for the creation, implementation, and analysis of data related to a computerized Raven-like test designed for the efficient assessment of fluid intelligence, named MatriKS. Emphasizing the utilization of knowledge space theory, the event delves into the benefits this theoretical framework provides in crafting and administering instruments that (a) ensure precise and efficient assessments of fluid intelligence, and (b) offer both quantitative and qualitative feedback potentially useful from a clinical perspective. The symposium presents the validation results obtained via multi-method techniques (classical test theory, item response theory, and knowledge space theory), for two distinct versions of the test catering to individuals aged 4 to 11 and 12 to 70. Moreover, it sheds light on the divergence and convergence outcomes when compared to traditional assessment instruments. In addition, some clinical implications are drawn from the analysis of the nature of incorrect responses to the test items. Finally, the perceived ease of use and attitude towards the computerized test are discussed for both the test-takers and the evaluators.

Discussant name:
Discussant surname:
Discussant affiliation:

FRIDAY 5 JULY
Session 9.1 SYMPOSIUM
Topic: Innovations in test development

531. Convergent and divergent validity of MatriKS: A new tool to assess fluid intelligence (Innovations in test development)

Alice Bacherini, Irene Pierluigi, Giulia Balboni
University of Perugia, Italy

Fluid intelligence (FI) represents the ability to think logically, process new information, learn, and solve problems in novel situations. FI deficits frequently occur in individuals with neurodevelopmental or psychiatric disorders. However, FI assessment is challenging because of the long time required by standard testing procedures. MatriKS, a new adaptive computer-based measure of IF, was developed based on Knowledge Space Theory. This contribution aims to investigate the convergent and divergent validity of MatriKS compared to traditional measures of FI (i.e., Raven's Matrices) and planning abilities (i.e., Tower of London, ToL), respectively. Based on the chronological age of the evaluated individual, a different version of MatriKS (4-11 or 12+ years old), Raven's Matrices (Colored, Standard, or Advanced), and ToL (4-13 or 14+ years old) were administered to 348 children and adolescents aged 4-18 and 102 adults aged 20-70 of the Italian general population. Pearson's correlation coefficients calculated between MatriKS and Raven's Matrices ranged from .46 to .82. They were statistically significantly higher than the corresponding divergent validity correlation coefficients between MatriKS and ToL (range: .08-.40), as compared using the Williams t-test for dependent overlapping correlation coefficients. These findings show that MatriKS is a cutting-edge instrument that provides a valid measure of FI in the considered context and is useful in clinical and research practice.

FRIDAY 5 JULY
Session 9.1 SYMPOSIUM
Topic: Innovations in test development

706. Enhancing the assessment of fluid intelligence with MatriKS: Insights from knowledge space theory and multi-method analysis techniques (Innovations in test development)

Debora de Chiusole, Ottavia M. Epifania, Andrea Brancaccio, Pasquale Anselmi, Luca Stefanutti
University of Padua

Knowledge space theory (KST) is a mathematical framework initially designed for efficient assessment and personalized learning. Its advantages in constructing and administering assessment instruments extend to both the quality of assessments and their outcomes. Indeed, accurate and efficient assessments are achieved, providing quantitative and qualitative feedback that is potentially valuable from a clinical perspective. While KST has traditionally been applied predominantly in knowledge assessment, recent extensions demonstrate its relevance to psychological and planning-skill assessment. This study presents the results of the first attempt to employ KST in assessing fluid intelligence. After presenting the construction and implementation phases of a new computerized Raven-like test, named MatriKS, the results of the Italian validation are shown. These results are provided for two distinct versions of MatriKS, adequate for individuals aged 4 to 11 (N = 632) and 12 to 70 (N = 838). For validation, a multi-method approach was adopted, incorporating classical test theory, item response theory, and KST. All models used in data analysis align with the overall construct validity and reliability of both versions. However, variations emerge in item selection and test dimensionality, warranting additional research on these aspects. In conclusion, we delve into several methodological and clinical/practical benefits associated with utilizing KST for constructing assessment tests and employing multi-method analysis techniques.

FRIDAY 5 JULY
Session 9.1 SYMPOSIUM
Topic: Innovations in test development

538. "Italian adaptation of the System Usability and Acceptance Model scale: application to MatriKS a new digital test for fluid intelligence assessment." (Innovations in test development)

Matilde Spinoso, Noemi Mazzoni, Matteo Orsoni, Mariagrazia Benassi, Sara Giovagnoli
Department of Psychology Renzo Canestrari, University of Bologna

The perceived ease of use and attitude towards new technologies are particularly relevant for the effectiveness and implementation of new digital neuropsychological tests. Nevertheless, to date there are no instruments specifically targeted for developmental age that encompass both the features of usability and acceptability. The present work aims to describe the psychometrics properties of the Usability and Attitude Scale (UAS), an Italian adaptation of the System Usability Scale (SUS; Brooke,1995) and Technology Acceptance Model Scale (TAM; Venkatesh & Bala, 2008; Davis, 1989). Furthermore, the study aimed at assessing the usability and attitude of MatriKS, a new computerized test for the evaluation of fluid intelligence. The UAS was administered to a sample of 1,239 participants aged 4 to 70 (47% males and 53% females, M=19.43, SD=16.86) of the general population. All participants completed different versions of MatriKS according to their age and then completed the UAS. Overall, UAS showed good content validity and internal consistency. Moreover, the results obtained using the Curved Grading Scale Method (CGS; Sauro & Lewis, 2012;2016), confirmed the positive reception of MatriKS, indicating that test-takers perceived the assessment as a positive experience. Specifically, the usability resulted "excellent" for the version tailored to the younger population, while the usability score of the other version resulted between "sufficient" and "good". Concerning the acceptability perceived by participants, the results obtained with the CGS method, confirm positive ratings like those found for usability. In conclusion, the UAS showed good psychometric properties and feedback provided by test-takers suggest that the MatriKS is promising and deserves further research.

FRIDAY 5 JULY
Session 9.1 SYMPOSIUM
Topic: Innovations in test development

556. **Noemi Mazzoni (University of Bologna), Matilde Spinoso (University of Bologna), Sara Giovagnoli (University of Bologna), Matteo Orsoni (University of Bologna), Sara Garofalo (University of Bologna), Mariagrazia Benassi (University of Bologna)**
Developmental trajectories of accuracy and type of errors in fluid intelligence assessment as detected by a new digital tool: MatriKS (Innovations in test development)

WEDNESDAY 3 JULY
Session 9.2 SYMPOSIUM
Topic: International assessment

75. Implementation of Teacher Performance Assessments Internationally

Jon Twing
University of Sydney
Janet Clinton,
University of Melbourne
Mark Grant
AITSL
Wayne Cotton
University of Sydney

Internationally, teacher excellence is seen as one way to improve education access for everyone, thereby providing more equity and opportunity for all learners. While many reform efforts in education have focused on improvements of learning, this symposium focuses on the use of teacher performance assessments, their application and implementation as a learning and policy tool internationally. The symposium is comprised of teacher education and teacher leader experts with many years of experience in teacher education across a variety of contexts and countries. Professor Grant will discuss the efforts of the Australian Institute for Teaching and School Leadership (AITSL) and discussed teacher preparation and performance assessment for Australia's education system, Australia's national accreditation system for initial teacher education programs—with a focus on teaching performance assessments (TPAs) and the evolution and challenges in its implementation. Professor Clinton will discuss the path to implementing TPAs from a provider and consortium lead perspective. Implementation of performance assessments for teachers as the responsibility of initial teacher education providers and as a point of contention—variable contexts, institutional philosophical stance, timelines, and programs were seen as potential impediments to a uniform implementation. Various approaches to the assessment lens, modes and implementation philosophy were also threats to validity across programs. Professor Clinton will discuss how some of these challenges were overcome. Professor Cotton will reflected on how a TPA was implemented specifically at the University of Sydney and the complicated nature of navigating consensus and change at a large and diverse institution with a proud heritage of supporting education and teacher education. Finally, Dr. Twing will discuss teacher education reform efforts in the US, India and Saudi Arabia with an eye to the research and implementation challenges.

Discussant name: Lisa
Discussant surname: Keller
Discussant affiliation: University of Massachusetts

WEDNESDAY 3 JULY
Session 9.2 SYMPOSIUM
Topic: International assessment

809. Teacher Preparation and Performance Assessment (International assessment)

Mark Grant
Australian Institute for Teaching and School Leadership (AITSL)

This Session outlines Australia's education system and the role AITSL plays in it as well as more details about the Australian national accreditation system for initial teacher education programs This Session will focus primarily on teaching performance assessments (TPAs), the challenges encountered in their implementation, and the roles they will play in the future of education in Australia and Australia's education reform efforts. As an Australian Government Agency, AITSL plays an important role in supporting consistent standards for teaching across Australia. AITSL's national frameworks are agreed at the national level, with education systems and sectors then taking on responsibility for implementation—providing a range of support including standards, frameworks, tools and resources. In this session, I will discuss The Australian Professional Standards for Teachers, or the Teacher Standards that underpin AITSL's national frameworks. They comprise seven Standards which outline what teachers should know and be able to do. The Standards were based on existing Australian and international standards, validated with 6,000 teachers and principals. The Standards provide a consistent, nationally accepted definition of what quality teaching looks like. Aligned to these Standards are 12 teaching performance assessments (TPAs). These TPAs are: reflective off classroom teaching practice; assess the Standards; have achievement criteria; demonstrate reliability of consistent scoring; and, include moderation processes that support consistent decision making against achievement criteria. Some of the implementation challenges to be discussed include: The number of TPAs could not be restricted; application for endorsement of a particular TPA could not be restricted; No time limits could be placed on the endorsement of TPAs; Participation in the trailing of TPAs could not be mandated; Standard setting or benchmarking across institutions could not be mandated or standardised. These challenges, mitigations and explanations will be discussed.

WEDNESDAY 3 JULY
Session 9.2 SYMPOSIUM
Topic: International assessment

810. Assessment for Graduate Teaching (AfGT) (International assessment)

Janet Clinton
University of Melbourne

This Session presents how a consortium of Universities in Australia with initial teacher preparation programs came together to build appropriate teacher performance assessments known as the Assessment for Graduate Teaching (AfGT). The AfGT was developed by teacher educators for teacher education and was in response to AITSL's mandate as discussed in another Session in this symposium. The AfGT was approved by AITSL's Expert Advisory Group in 2019 and was reviewed as: "A well thought out and thorough TPA that demonstrates a valid and reliable method for assessing whether a teacher's performance meets the Australian Professional Standards for Teachers at the Graduate Teacher level". Four basic attributes were developed for the AfGT: planning documentation, observations on and evidence of practice, annotated samples of candidate work, and teaching scenarios. The AfGT is a holistic, measurable, research informed national summative assessment to demonstrate classroom readiness at point of graduation. It calls upon the candidate's knowledge and skill in making situational judgements about scenarios which they may face as beginning teachers and, as such, represents much more than content knowledge or basic pedagogy. To ensure the fidelity or trustworthiness, it is not possible to change or modify the AfGT and all elements must be completed and passed. The AfGT is only one requirement that a candidate must pass in addition to obtaining the professional experience during placement. The AfGT is an assessment task in a student's coursework program and requires consent to video recording and using student work samples. The 'impact' of the AfGT does not sit well philosophically with much of our institutional pedagogy. Having to provide evidence that what we teach really that impacts a child's learning is difficult to show in an institution context. That is the same for assessment, i.e., it does not fit philosophically to assess children in the way of the standards and the way AfGT requires. These challenges and others will be addressed in this session.

WEDNESDAY 3 JULY
Session 9.2 SYMPOSIUM
Topic: International assessment

811. A TPA called the AfGT with AI & ML (International assessment)

Wayne Cotton
University of Sydney

How the AfGT was implemented at the University of Sydney will be explained. The AfG is indeed an assessment, but it is also an ongoing research project. The assessment portion itself contains 3,000 words, covers the four standards elements (planning, teaching, assessing, and reflecting), is taken during a pre-service teacher's final internship and is marked by university staff. Planning for teaching and learning is focused on a candidate's capacity to understand the context of their placement, their planning for student learning in relation to the specific goals of a sequence of lessons, and the ways in which they will judge their impact on school student learning. Analysing teaching practice focuses the pre-service teacher's capacity to understand the implications of pedagogical practice on student learning. The AfGT gathers robust evidence for reflection using mentor reports, school student work samples, peer reports (if possible), and videos of key pedagogical segments (two 6-10min clips). Assessing for impact on student learning is focused on the pre-service teacher's capacity to implement a targeted summative assessment task and their ability to understand the extent to which students in the classroom have been able to achieve the learning goals. Expanding practice (situational judgement) is completed online where pre-service teachers address standards like "Demonstrate broad knowledge of, understanding of and respect for Aboriginal and Torres Strait Islander histories, cultures and languages." This section of the AfGT is scenario based and is supported with strong measurement rubrics. The University of Sydney has trained stakeholders in how to record videos and meet the ethical requirements of the of the AfGT. We have introduced data literacy units as part of instruction due to the Standards measured by AfGT. To improve efficiency, lower expense, and time burdens, we have explored the use of AI/ML (e.g., automation) to mark elements of the AfGT. We are investigating the predictive validity of the AfGT and have established strict protocols to help ensure video privacy for our candidates. This Session will show anyone who wants to implement teacher performance assessments how it can be done or was done in at least the context of the University of Sydney.

WEDNESDAY 3 JULY
Session 9.2 SYMPOSIUM
Topic: International assessment

812. The Similarity and Differences in Teacher Performance Assessments Internationally (International assessment)

Jon Twing
University of Sydney

In this session, Dr. Twing will compare and contrast assessment, educational reform efforts and teacher professional development world-wide. In his more than 40 years supporting educational reform efforts world-wide, he is often surprised in the similarities in challenges resulting from both circumstances and implementation of best practices regardless of local jurisdiction. Specifically, Dr. Twing will review teacher education reform efforts in the United States, India, Saudi Arabia, and Australia. This review will include relevant standards and how they came about; ministry policy and how it guides, supports or otherwise exacerbates education reform efforts; the impact capacity and infrastructure (or lack thereof) impacts reform efforts as they relate to teacher education., the role performance assessments plan in educational reform and why they seem so controversial and the role professional development, specifically for teacher education. Finally, a summary comparison on the similarities and differences across the symposia sessions (including his own) as well as what he has seen and discussed around the world will also be presented.

WEDNESDAY 3 JULY
Session 9.3
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

149. Psychological assessment in Latin American countries: Perspectives and Challenges

Solange Muglia Wechsler
Pontificial Catholic University of Campinas-Brazil

The history of psychological assessment in Latin American countries is little known. Although South and Central American nations were influenced by the development of psychological tests in Europe and North America, there were distinct developments in each of these cultures, which impacts the current conditions of psychological assessment until today. An overview of the situation in many of these countries will be presented. Among these countries, Brazil has the better development in psychological tests, as there is a national system to evaluate the quality of tests being used, and more national tests are being constructed to attend to the population characteristics. Other countries such as Chile and Peru are also striving to achieve better quality of psychological services in their regions, thus stressing the need to establish formal regulations for the use of tests and implement the creation and adaptation of tests validated to their population. An international study obtained with psychologists from different Latin American countries which indicated their attitudes toward testing, difficulties, and challenges will be discussed.

Discussant name: José
Discussant surname: Muniz Fernandez
Discussant affiliation: Universidad Neblija-Spain

WEDNESDAY 3 JULY
Session 9.3
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

414. Innovative Development and Validation of Tale Me More®, a Multicultural Professional Profiling Model: Integrating Theory and Practice

Céline Jouffray, Eric Duvergey
Talent Tale - France
Martin Storme
IESEG - France

In collaboration with a talent management consulting firm, we developed a questionnaire to explore an individual's professional profile in terms of their skills and behaviors, free from any judgment. We based our approach on differential and situationist personality assessment methods and combined theoretical approaches (Big Five, Self-Determination Theory…) and empirical insights, based on thousands of interviews conducted within the firm. This led us to a model comprising 15 personality traits and a forced-choice questionnaire of105 pairs of statements – 14 statements per trait. The questions were constructed in French, and culturally adapted into 6 languages. A sensitivity study of the items was conducted for each language to enhance cultural adaptation across samples ranging from 65 to 1185 participants. Items were adjusted and pretested, in French, on a sample of 1256 employees (766 men, 490 women) aged 20 to 64 (average 40 years), with a higher level of education and several years of professional experience. A Thurstonian model was estimated using thurstonianIRT (Bürkner, 2019). To ensure the fairness of the evaluation process we examined differences in latent trait scores based on respondents' demographic characteristics. The complexity of the model prevented the software from estimating the Thurstonian model, but the Bayesian framework provided by Stan overcame this complexity, confirming the stability of Bayesian estimation. Latent scores were estimated for subsets, showing empirical reliability exceeding .50 for all traits, with some being assessed more reliably than others. A series of multiple regression analyses was conducted with the latent score of each trait as the dependent variable, and age, gender, and education level as independent variables. The analyses revealed differences in scores highlighting the recommendation to standardize scores. A test-retest reliability study and concurrent validity with an adapted version of the IPIP are underway.

WEDNESDAY 3 JULY
Session 9.3
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

594. Innovative Development and Validation of Tale Me More®, a Multicultural Professional Profiling Model: Integrating Theory and Practice

Wendy Magee
Temple University Philadelphia USA
Jonathan Pool
Anglia Ruskin University Cambridge United Kingdom

Disorders of Consciousness (DoC) following acquired brain injury are characterised by poor wakefulness and complex impairments in the motor, cognitive and sensory domains. Assessing awareness is essential to determine appropriate care. However, measures need to meet both neurodevelopmental changes and acquired impairments. Language-based measures are not optimal due to developmental differences complicated by cognitive impairment. Yet, internationally, validated assessments of awareness are lacking and children with DOC remain one of the most under-researched populations. Based on an existing validated adult scale, the Music therapy Sensory Instrument for Cognition, Consciousness and Awareness (MuSICCA) is a new music-based measure for assessing sensory responsiveness and awareness in paediatric DOC. We aimed to establish face validity of the MuSICCA as part of a larger international validation study. Twenty participants with significant experience of DoC were recruited within the UK: 15 professionals (10 music therapists and 5 allied health professionals) and 5 family members of children with DoC. Quantitative data were collected to establish agreement/disagreement with qualitative statements about the tool's strengths and weaknesses and analysed using descriptive statistics and thematic analysis. 100% of the participants agreed that the MuSICCA offers a clinical tool for assessing DoC in children. Findings reveal the MuSICCA design offers a comprehensive sensory assessment that is rigorous through being evidence-based. While its utility is multi-faceted, the MuSICCA requires some further refinement and the need for training reduces accessibility. The results suggest face validity for the MuSICCA. The tool benefits from caregiver involvement and multidisciplinary perspectives enhancing how the findings can be meaningful to all stakeholders. Psychometric validation is in progress.

WEDNESDAY 3 JULY
Session 9.3
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

600. Examining the Predictive Validity of the Romanian Adaptation of the Conners-3 Short Form for ADHD Diagnosis

Șerban Zanfirescu Zanfirescu, Dragoș Iliescu
University of Bucharest
Andreea Butucescu
andreea.butucescu@unibuc.ro

The Conners-3 is an established psychological test for assessing attention deficit/hyperactivity disorder (ADHD) symptoms in children and adolescents. It relies on reports from parents, teachers, and self-reports, aligning with diagnostic criteria for ADHD and other conduct disorders. Following international testing standards, the Conners-3 has been adapted into various languages, including Romanian. However, discrepancies in diagnostic outcomes have emerged, particularly with the Romanian short form. This study investigates these variances by analyzing a subset of items from the Conners-3 for their predictive accuracy in diagnosing ADHD. The present study encompassed 1081 participants (Nboys = 539, 49.86%; Ngirls = 542, 50.13% girls; Mage = 12.58 years, SD = 3.65). Each participant was evaluated using the Conners-3, and data also contain the ADHD diagnosis. Based on the screening questionnaire administered, participants in the standardization sample (N = 1028) are assumed to not have the diagnosis. A number of 53 participants were added to this standardization sample, who were diagnosed with ADHD based on psychiatric interviews. To determine the most predictive items for the Romanian Conners-3 short form, a series of analytical procedures were employed. These include decision trees, random forest regression, exploratory and confirmatory factor analyses, logistic and LASSO regressions, and item characteristic curve analysis from item response theory (IRT) models. Several short form scales of the Conners-3 were thus developed and assessed for diagnostic accuracy. The implications of this study are significant, enabling both researchers and practitioners to more effectively use the Conners-3 short form for screening individuals for ADHD and related disorders. From a psychometric point of view, the study underscores the importance of adaptation not only in the wording but also the scoring of items and scales, for increased cross-cultural (criterion) validity.

WEDNESDAY 3 JULY
Session 9.3
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

698. Validity evidence of the Pornography Consumption Inventory: Relationship with subjective and objective sexual arousal measures

Oscar Cervilla Saez, Ana Álvarez-Muelas, Laura E. Muñoz-García, Pablo Mangas, Gracia M. Sánchez-Pérez, Reina Granados, Juan Carlos Sierra
Mind, Brain and Behavior Research Center (CIMCYC), University of Granada

Introduction. The Pornography Consumption Inventory (PCI) is a scale that assesses pornography consumption in three dimensions: Emotional Avoidance (EA), Sexual Curiosity (SC), and Excitement seeking and sexual pleasure (ES). There is little evidence for the validity of this scale. Objective. To provide validity evidence of the Pornography Consumption Inventory by relating measures of its three dimensions to different manifestations of sexual arousal (i.e., propensity for sexual excitation/inhibition, ratings of self-reported sexual arousal and genital sensations, and genital response). Method. In 73 heterosexual men and 70 heterosexual women, aged 18-30 years, we first assessed propensity for sexual excitation and inhibition and, then, recorded their self-reported sexual arousal and genital response (i.e., penile erection or vaginal pulse amplitude) to viewing neutral and sexually explicit films of the preferred gender. Pearson correlations were obtained between PCI dimensions and sexual arousal measures. Results. In men and women, EA was positively related to the propensity for sexual excitation ($r = .34$, $p < .01$; $r = .32$, $p < .01$; respectively for men and women) and the ratings of self-reported genital sensations ($r = .25$, $p < .05$; $r = .33$, $p < .01$), whereas SC was positively related to propensity for sexual excitation ($r = .25$, $p < .05$; $r = .31$, $p < .01$). In women, moreover, SC correlated positively with the ratings of self- reported sexual arousal ($r = .17$, $p < .05$) and negatively with sexual inhibition due to the threat of the consequences of sexual activity ($r = -.30$, $p < .05$). Finally, ES, in both men and women, was positively related to the ratings of genital sensations ($r = .24$, $p < .05$; $r = .28$, $p < .05$) and, in men, to the propensity for sexual excitation ($r = .45$, $p < .001$). Conclusions. The three dimensions of the Pornography Consumption Inventory are associated with different manifestations of sexual arousal, thus providing validity evidence.

WEDNESDAY 3 JULY
Session 9.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

246. Evaluating an AI-based method for predicting language DIF in high-stakes, cross-language science assessments

Joshua McGrane,
The University of Melbourne
Heather Kayton
The University of Oxford
Zhonghua Zhang
The University of Melbourne

Despite rigorous undertakings by test-designers to ensure comparability of cross-language assessments, studies often expose the differential performance of items across language versions. Typically, such studies use Differential Item Functioning (DIF) analysis to identify problematic items, followed by qualitative review of the item content across languages using expert judges. However, these review procedures are often ineffective in explaining the occurrences of DIF and are very time and resource intensive. While researchers agree that linguistic features of items contribute to comparability across languages, it remains extremely difficult to isolate potential language effects on item functioning. Fortunately, advances in Natural Language Processing (NLP) present opportunities that make it possible to evaluate linguistic features of items, and particularly different aspects of textual complexity, in efficient and comprehensive ways. This study applied techniques from psychometrics, machine learning and NLP to evaluate a method to predict language effects on item comparability in a high-stakes, international cross-language assessment. An exploratory model was built using Random Forest Regression to investigate the extent to which differences in language features of items predict DIF magnitude between English and French/Spanish language versions of six of the International Baccalaureate's Diploma Programme science assessments. The model showed that textual complexity features explained between 11% and 13% of the DIF variance. The combination of conventional psychometric DIF modelling with advances in AI modelling provides an approach that is efficient, scalable and, moreover, may be used to predict potential DIF items prior to test administration. This is important in high-stakes assessments where piloting prior to administration is often not possible. We believe the approach shows promise for future refinements, developments, and applications.

WEDNESDAY 3 JULY
Session 9.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

307. **Asymmetrical Item Response Models**

Jorge Bazán
University of São Paulo

A new wave of asymmetrical Item Response Models has emerged in psychometric literature, revolutionizing our understanding of latent ability manifestation through novel shapes in Item Characteristic Curves (ICC). These models, prompted by an observed surplus of either correct or incorrect responses, underscore their significance in capturing response patterns unaccounted for by traditional IRT models—especially in scenarios where correctly answered items pose distinct difficulties. In this study, we introduce two pioneering models: the Skew Item Response Model and the Logistic Positive Exponent Model, alongside their respective reversed counterparts. We elucidate the Bayesian estimation process for these models, offering a robust methodology for deriving accurate parameter estimates. Through practical examples, we illustrate the enhanced flexibility of IRT models with an additional parameter, showcasing their efficacy in representing intricate ICC patterns observed in educational and psychological assessments. This research not only contributes to the refinement of psychometric tools but also emphasizes the relevance of embracing nuanced models that can better capture the complexities inherent in the diverse response patterns exhibited across various applications in education and psychology.

WEDNESDAY 3 JULY
Session 9.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

395. Cross-Cultural Validation: English version of the Problematic Use of Social Networking Sites Questionnaire

Covadonga González-Nuevo
University of Burgos
Álvaro Postigo, Jaime García-Fernández
University of Oviedo
David Jhonston, Chloe Ryding
Nottingham Trent University
Marcelino Cuesta
University of Oviedo
Daria Joanna Kuss
Nottingham Trent University

Introduction: The Problematic Use of Social Networking Sites (SNS) questionnaire (PUS) not only assessed the addictive consequences of SNS use but also negative social comparisons. However, this questionnaire had not been validated in English, thereby preventing cross-cultural comparisons. Therefore, the primary objective of this study was to validate the PUS questionnaire in English. Method: The sample comprised 654 participants from the UK, aged between 18 and 78 years (M = 33.69; SD = 13.17). Time spent using SNS, problematic use of SNS (using the PUS), and emotional distress (using the Hospital Anxiety and Depression Scale, HADS) were assessed. We conducted a Confirmatory Factor Analysis (CFA) with the UK sample to confirm the two-factor structure of the PUS, comprising the subscales of Negative Comparison and Addictive Consequences. Results: CFA showed an adequate fit to the two-factor structure (CFI = .96; RMSEA = .07), and the internal consistency of the Negative Comparison and Addictive Consequences subscales was excellent (α = .94 and α = .91, respectively). A positive correlation was obtained between Negative Comparison with anxiety (r = .57, p < .001) and depression (r = .43, p < .001) as well as between Addictive Consequences with anxiety (r = .42, p < .001) and depression (r = .29, p < .001), as expected in literature. Conclusion: The PUS questionnaire demonstrated excellent psychometric properties in the UK sample, hence it can be used to assess both addictive and comparative use of SNS by users in the UK and conduct cross-cultural studies.

WEDNESDAY 3 JULY
Session 9.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

488. Psychometric Investigation of Fairness: Role of Test Content Language and Student's Language

Jordan Southcott
Multi-Health Systems

Theoretical Framework: Assessments with no language within item content theoretically increase equity amongst different primary language groups. Differential Test Functioning (DTF) is a psychometric method to assess measurement invariance across groups. An ANCOVA assesses group differences while holding constant demographic variables that need to be controlled in the model. Objectives: The focus of this analysis is to assess the fairness of an assessment with items that contain no language, the Naglieri General Ability Tests, for students with different primary languages. Sample: The sample includes students in the US enrolled in Kindergarten through Grade 5 (N = 6,000). Two groups were used for analyses: students who primarily speak English (n=2,914) and students who primarily speak other languages (n=445). Methodology: The study comprised 3 tests: Naglieri–Verbal, Naglieri–Nonverbal and Naglieri–Quantitative. Signed DTF (sDTF), Unsigned DTF (uDTF), and expected test score standardized differences (ETSSD) were used as measures of DTF. ANCOVAs were used to compare mean scores between groups, using gender and pre- vs. post-pandemic outbreak as covariates. Results: Minimal DTF was detected. All sDTF and uDTF were less than 4% and ETSSD less than |.10|. Mean differences for Naglieri–Verbal and Nonverbal were not significant, with negligible effect sizes. Naglieri–Quantitative showed significant differences between students who primarily speak English and those with other language backgrounds ($p < .001$), but effect size was small (Cohen's $d = 0.26$). Implications: The absence of measurement bias and the small differences between linguistic groups highlight the equitable nature of an assessment that minimizes reliance on language in item content. These findings are important, as grasp of culturally dominant languages can confer unfair advantage to test takers and blunt the ability to detect giftedness in other language populations.

WEDNESDAY 3 JULY
Session 9.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

724. Ethical principles in Artificial Intelligence telerehabilitation for neurodevelopmental disorders: Development of a questionnaire for survey research

Aurora Castellani
University of Perugia
Mariagrazia Benassi
University of Bologna
Giulia Balboni
University of Perugia

Recently, telerehabilitation techniques for neurodevelopmental disorders (NDD) have been implemented, automating the rules for setting the intervention protocol using Artificial Intelligence (AI). Although these methods could have advantages, such as personalization of treatment and self-adaptivity, it is unclear how ethical principles can be applied to these interventions. This study describes the development of a questionnaire investigating the attitudes, feelings, and expectations towards the ethics of AI telerehabilitation programs for children with NDD. Recipients will be (1) clinicians, (2) software developers, (3) ethicists, and (4) caregivers utilizing these systems. We reviewed the scientific literature on ethics and AI for this aim and used linguistic content analysis to identify the most relevant ethical principles. Then, we developed a preliminary version of the questionnaire. This version investigates stakeholders' knowledge of the ethics of AI with children as users and the importance assigned (Likert Scale of 1 to 4 points) to specific ethical principles with their examples of application in the use of AI in telerehabilitation. Now, we are ready for field testing. Five participants will be recruited for each of the four stakeholder groups. They will be asked to evaluate each item for (1) clarity and comprehensibility; (2) usefulness of provided examples; (3) relevance of each ethical principle; and (4) completeness to ensure that the questionnaire effectively encompasses a wide range of ethical principles linked to AI in NDD telerehabilitation; (5) appropriateness of response scales. We will develop and revise a second questionnaire version based on stakeholder feedback. The final version will be used in a future research survey. The aim will be to develop an ethical best practices checklist for designers and clinicians to assess if a telerehabilitation program aligns with ethical principles and the priority provided by all stakeholders.

WEDNESDAY 3 JULY
Session 9.5 SYMPOSIUM
Topic: Psychometric modeling

454. Forced-Choice Measurement - Investigating Response Processes

Susanne Frick
TU Dortmund University

In Forced-Choice (FC) questionnaires, respondents indicate their relative preference for items within blocks. The comparative response format allows to reduce faking but elicits a response process that is not yet well understood. The contributions in this symposium highlight different aspects of the FC response process and how it is shaped by FC test design. Anna Brown presents a response model for faking on an item-level in FC questionnaires and validates it in an empirical study with software engineers. Susanne Frick investigates how desirability matching relates to indicators of response editing in an analysis across six datasets with manipulated high and low stakes conditions. Nigel Guenole presents the trifactor change model for understanding how item properties differ between high- and low-stakes conditions and between rating scale and FC responses. Rebekka Kupffer presents a study in which careless responding was manipulated experimentally and its effect on careless responding indices was investigated. The contributions in this symposium illuminate the FC response process from different perspectives and can inform FC test construction.

Discussant name:
Discussant surname:
Discussant affiliation:

WEDNESDAY 3 JULY
Session 9.5 SYMPOSIUM
Topic: Psychometric modeling

456. Modelling 'intermittent faking' on forced-choice questionnaires (Psychometric modeling)

Anna Brown
University of Kent

Personality is almost exclusively assessed using self-report questionnaires, which are open to respondents creating the best impression (aka 'faking good'). To counteract faking, the use of 'forced-choice' questionnaires has been popular since appropriate item response models became available to scale them, for example Thurstonian IRT models (Brown & Maydeu-Olivares, 2011, 2018). Forced-choice questionnaires are particularly effective at preventing faking when all choice alternatives appear equally desirable (Cao & Drasgow, 2019). However, to evaluate the effectiveness of matching alternatives on desirability (and potentially correcting for the lack thereof), methods for detecting and measuring faking in forced-choice questionnaires are needed. OBJECTIVES This talk will present a response model for faking in forced-choice questionnaires, akin to the recent model of 'intermittent' faking for rating scales (Brown & Böckenholt, 2022). METHODOLOGY I consider forced-choice responses of every test taker as a potential mixture of 'real' (or retrieved) choices, and 'ideal' (manipulated) choices, with each response type characterized by its own distribution and factor structure. I model the data using two-level factor mixture models, with item choices nested within the person. This approach allows modelling faking as person-by-item interactions, taking to account both item properties (for example, desirability matching) and personal attributes relevant to faking. ILLUSTRATION The approach is illustrated with an operational study of software engineers (N=2039), in which stakes were manipulated. A bespoke questionnaire measuring 5 job-related traits with 20 forced-choice blocks of 3 statements was developed for this study. I demonstrate that every person-choice can be probabilistically classified as 'real' or 'ideal', measuring the person's grade of membership in both profiles. I use relevant auxiliary variables (e.g. response latencies) to validate the response process.

WEDNESDAY 3 JULY
Session 9.5 SYMPOSIUM
Topic: Psychometric modeling

470. Trifactor Change Models for Likert and Force Choice Questionnaires (Psychometric modeling)

Nigel Guenole
Goldsmiths, University of London
Anna Brown
University of Kent

Understanding how personality item properties differ across low and high stakes conditions is of great interest to practitioners. One method for studying this question is the simulated faking design where respondents answer under honest and faking instruction sets. Studies so far have commonly used observed level scores, or alternatively, where latent models are used, have not incorporated the invariance constraints and other modeling features necessary for robust conclusions about item performance under each condition. Guenole, Brown and Lim (2023) introduced the Trifactor Change model for this situation, but so far, the model has only been implemented with measures using two scales and with Likert items. OBJECTIVES This talk will discuss the trifactor model applied to a six-dimensional maladaptive personality questionnaire, the G60, showing Likert results and generalizing to Forced Choice. METHODOLOGY In the Trifactor Change model, we model three sources of variance: substantive factors (here, the six DSM-5 maladaptive personality traits), condition related common factors (honest versus faked responses), specific factors to account for repeated exposure (as correlated residuals), and a latent mean structure to allow for the faking effect. If invariance sufficiently holds, we may interpret the latent mean difference as a condition related difference in latent means due to faking. ILLUSTRATION 516 participants answered the G60 Derailers questionnaire (Guenole, 2015; Guenole, Brown, & Lim, 2018) under honest and instructed faking conditions. In the Likert sample, we fitted the trifactor model of Guenole et al. (2023) to 120 items (60 items at two time-points) with sufficient invariance to interpret latent mean change. We will compare the invariance of the Likert items to the Forced Choice version of the G60/Derailers measure using the same sample.

WEDNESDAY 3 JULY
Session 9.5 SYMPOSIUM
Topic: Psychometric modeling

463. Careless Responding in Multidimensional Forced-Choice Questionnaires: What Does it Look Like and how can it be Detected? (Identifying biases by qualitative or quantitative methods)

Rebekka Kupffer
University of Kaiserslautern-Landau
Susanne Frick
TU Dortmund University
Eunike Wetzel
University of Kaiserslautern-Landau

Careless responding is a response behavior in self-report questionnaires characterized by selecting response options without considering the (whole) item content. Currently, little is known about how careless responding manifests itself in multidimensional forced-choice (MFC) questionnaires. In a laboratory study (N = 430), we manipulated careless responding and compared its manifestation on eleven careless responding measures. Participants were assigned to one of three conditions: a control condition (n_1 = 140) and two careless responding conditions in which we either instructed the participants to respond carelessly (n_2 = 140) or distracted them to induce careless responding (n_3 = 150). The survey consisted of five inventories assessing personality traits in the MFC format, and concluded with an oral interview with questions about the response process. To detect careless responding, we included self-report items on data quality and instructed response triplets. We also calculated post-hoc measures, including analyses of response times and rank order patterns, as well as consistency and outlier analyses. We found significant differences between the control and the instructed careless responding condition on all of the indices except for the response times. However, the control and the distracted responding conditions did not differ significantly. In an exploratory comparison of the instructed and the distracted conditions, we found small to large differences on most indices. In the interview, commonly reported strategies for being listless, tired, or unmotivated were reading the items superficially, thinking less about the item content, choosing a random order, or copying the presented order. Practical implications for identifying careless responding in MFC questionnaires are discussed.

WEDNESDAY 3 JULY
Session 9.5 SYMPOSIUM
Topic: Psychometric modeling

455. Using Process Data to Understand Response Processes Underlying Faking in Questionnaires (Psychometric modeling)

Susanne Frick
TU Dortmund University
Miriam Fuechtenhans, Anna Brown
University of Kent

Background Impression management (aka Faking) on self-report questionnaires is a concern in high-stakes assessments. The forced-choice (FC) format has been proposed to overcome faking. However, faking resistance depends on the item desirability and on the desirability matching. Process data, such as response times or changes made to initial response, can help understand the process of faking. Objective The objective of this study is to investigate how item and person characteristics related to faking manifest in response editing in questionnaires. Samples We conducted a re-analysis of 6 datasets, all of which represent responses to RS and/or FC questionnaires under controlled conditions (either low or high assessment stakes or both, and known item desirability), with process data such as response latency, number of clicks and others. Item desirabilities were rated on a rating scale by a separate sample. Methodology To each type of outcome variable (response time, clicks and desirable responses), we fitted models that account both for the variance due to items or blocks and persons. They can be equivalently seen as item response theory (IRT) models or cross-classified multilevel models. All models were of the Rasch-type, thus, we assume that all items are equally discriminating. Block or item level predictors were derived from the desirability ratings. Results We found shorter response times for FC blocks of items that were less well matched in desirability, although significant in only one study. The effect of ambiguous items differed between RS and FC. However, most interactions of the item covariates with stakes were not significant. Implications From a psychometric perspective, this study can inform further psychometric developments for the analysing of process data. From a practical perspective, the results of this research can inform the development of fake-resistant assessments and facilitate the evaluation of the impact of faking on current assessments.

WEDNESDAY 3 JULY
Session 9.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research/ Innovations in test development

69. **Future of AI ML in Software Testing**

Shreya Asthana
Red Hat

Artificial Intelligence (AI) is making a significant impact in multiple areas like medical, military, industrial, domestic, law, arts. But AI ML transformed software testing in ways that could not have been dreamt of a decade ago. The testing community is turning to AI to fill the gap as AI is able to check the code for bugs and errors without any human intervention and in a much faster way than humans. As AI becomes more mainstream, there are likely to be entirely new career fields that have not yet been invented. I aim to recognize the impact of AI technologies on various software testing activities in the STLC, explain some of the biggest challenges software testers face while applying AI to testing and how self healing will take care of your script and save script maintenance time. Audience also get to know some key contributions of AI in the future to the domain of software testing. Key Takeaways: how AI and ML can be used to improve your testing processes, leverage AI-based security tools, and implement risk-based methods such as risk-based testing that can leverage big-data insights. How to make the current automated tests more resilient and less brittle. How self healing observe changes in the application and start learning the pattern of changes and then can identify a change at runtime without you having to do anything.

WEDNESDAY 3 JULY
Session 9.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research/ Innovations in test development

87. **Innovations in test development**

Emeric Kubiak, Simon Baron, Tales Marra
AssessFirst

Objective. While Big Five are related to job-performance, personality assessments as still viewed unfavourably at times. Fortunately, new studies show that gamifying the assessment with images proves valuable for evaluating personality (Hilliard et al., 2022) while improving user engagement (Efremova et al., 2022). Also, the integration of images serves as a strategic approach to reduce textual content, thereby facilitating cultural adaptation. The goal of this study is then to: (1) develop an image-based, forced-choice, five-minutes assessment measuring BFI-2 and facets of Humility (Denissen et al., 2022), (2) discuss the use of machine learning to enhance content-validity, (3) examine it's psychometric properties. Method and Results. Study 1 outlines the development of our assessment across two samples (N=2,989; N=4,457), with participants completing image-based tasks alongside the BFI-2 questionnaire, including additional humility items. Study 2 illustrates leveraging DeBERTa (He et al., 2021) for content validity, demonstrating remarkable accuracy (.92), precision (.91), recall (.92), and F1-score (.91) across traits. Study 3 inspects the psychometric properties of the assessments on a sample of N=4,457, revealing satisfactory validity metrics (mean convergent validity of .70 with the BFI-2, mean RMSEA of .01, item-dimension saturation ranging from 46 ≤ r ≤ .58, and consistent inter-dimension correlation with BFI-2 = .66, p < 2.2e-16), reliable indices (mean McDonald's Omega of .76), excellent discrimination (Ferguson's > .9), and negligible gender effect (mean Cohen's d = .08). The impact on certain facets regarding gender aligns with prior research (e.g., Kajoniusa & Johnson, 2018). Conclusion. Our study introduces a groundbreaking approach in the field by employing image-based assessments to measure personality traits effectively. Practically, it suggests a more engaging and culturally adaptable method for companies to enhance their hiring processes.

WEDNESDAY 3 JULY
Session 9.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research/ Innovations in test development

172. Ambiguity in Educational Assessments: The Impact of LLM AI on Source Preprocessing in the Translation Workflow

Ken Clark, Gustavo Ribone
Responsive Translation / United States of America

Theoretical Framework: This study, motivated by the need for tools and methods to ensure clarity and accuracy in translation, specifically of educational assessments, draws on Computational Linguistics and Natural Language Processing principles from "Speech and Language Processing" by Jurafsky and Martin. The book's multimodal framework integrates Semantic Role Labeling to interpret word relationships based on context, maintain referential clarity, and ensure logical text flow. This study is predicated on the premise that Language Model algorithms, specifically the latest GPT-4, can identify ambiguities in educational assessment documents. Objectives: To assess the efficacy of LLMs, particularly GPT-4, in detecting and clarifying ambiguities in educational assessments, contributing to the translation quality improvement. Methodology: Our methodology employs a structured, multi-phase approach involving a Linguist specialized in the educational field and a Prompt Designer: • Phase 1: Pilot testing with two standardized educational assessments to develop and refine GPT-4 prompts for ambiguity detection. • Phase 2: Scalability testing with a broader set of documents to assess the refined prompts' effectiveness in handling diverse subjects and complexities. • Phase 3: Comprehensive analysis and documentation. Results: We aim to uncover the precision of GPT-4 in identifying ambiguities and to assess the effectiveness of prompts across a variety of educational documents. This will help establish benchmarks for LLM performance in linguistic analysis. Implications: The study has two implications. Firstly, it offers a pragmatic evaluation of AI's usefulness in linguistic analysis, underscoring its practicality and reliability in the educational assessment industry. Secondly, it demonstrates how AI can help improve the pre-processing stage in translating educational documents, thereby reducing the risk of misinterpretations in the translation workflow.

WEDNESDAY 3 JULY
Session 9.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research/ Innovations in test development

518. Establishing Validity Arguments in Automated Scoring Contexts: A Roadmap

Hillary Michaels, Gavan OShea, David Dorsey
Human Resources Research Organization

In recent years, assessment professionals have articulated an approach to establishing validity that many find more integrated, rigorous, and informative than the commonly evoked tripartite (content, construct, criterion-related) validity framework. This perspective, often referred to as the 'validity argument' approach, is not a radical departure from the tripartite view. Rather, it emphasizes the centrality of establishing clear, well-defined, testable arguments about what an assessment is measuring, how test scores should be interpreted, and why an assessment is useful and relevant in a given context. In practice, this often involves articulating and then rigorously testing if-then 'argument chains' mapping each element needed to establish a principled foundation for assessment scores' intended use. The advent of automated scoring, driven by continually evolving advances in large language models (LLMs) and other artificial intelligence (AI) variants, can leave assessment professions grappling for effective ways to map the validity of these innovative approaches. In our view, the validity argument approach is well-suited to this challenge. Our presentation will walk the audience through an 8-step validity argument roadmap we developed specifically for automated scoring contexts (Dorsey, Michaels, & Ferrara, In press)—including targeting assessment constructs, intended score uses, test blueprints, item types, developing AI scoring models, administration procedures, computational and other psychometric analyses, and reporting. At each step, we will provide practical examples to facilitate attendees' application of the roadmap through their own work. For example, the roadmap element focused on model testing will include strategies to evaluate training data representativeness, scoring outcomes in terms of fairness, and 'stress tests' and longitudinal drift analyses that ensure the scoring model does not become less accurate or relevant over time.

WEDNESDAY 3 JULY
Session 9.6
Topic: Artificial Intelligence in testing, psychological assessment and survey research/ Innovations in test development

764. The Impact of DIF on Item Selection of Machine Learning Algorithms

Weldon Zane Smith, HyeSun Lee
California State University Channel Islands

The use of Machine Learning (ML) algorithms in daily life has increased greatly in recent years. While much attention has been given to algorithms imitating human learning and intelligent behaviors (e.g., ChatGPT), ML has also been applied in the testing industry. Much like how Computer Adaptive Testing has been employed to enhance the efficiency of testing, ML algorithms have also been applied to reduce the number of items on tests for more efficient administration that still retains high levels of accuracy. For example, Zheng et al. (2020) reduced a 173-item juvenile delinquency risk assessment down to 3-6 items using a tree-based ML algorithm (random forest). Similarly, Brodey et al. (2019) reduced a psychosis screening test from 124 items to 21. However, little investigation has been done on fairness of ML algorithms and the adverse impact of decisions made using ML algorithms. The reduction of tests from many items into few may lead to even greater impact of items exhibiting DIF, should those items be chosen for inclusion by the algorithm. The current study aimed to investigate what factors lead ML algorithms to include items exhibiting DIF, and how inclusion of those items impacts selection ratios for marginalized groups. implemented Monte Carlo simulations across 48 fully crossed conditions where magnitude of DIF (0.4, 0.8), number of items exhibiting DIF (10%, 20%), balance of majority/minority group membership (1:1, 4:1), and type of items exhibiting DIF (low/med/high difficulty, low/med/high discrimination) were manipulated. Results showed that ML algorithms tended to select items exhibiting DIF, especially when present in highly difficult or discriminating items and when the magnitude of DIF was large. The inclusion of items exhibiting DIF negatively impacted the selection of minority groups, sometimes resulting in as few as 3% being selected. Future directions were discussed, including how detection of DIF can be implemented within ML algorithms.

WEDNESDAY 3 JULY
Session 9.7
Topic: Identifying biases by qualitative or quantitative methods/ Translation of tests, psychological assessment instruments and survey questionnaire

97. Anchoring Vignettes: A Useful Tool to Measure and Correct for Cultural Bias in Parent Reports on Their Child's Mental Health?

Ronja Runge, Renate Soellner
University of Hildesheim, Germany

Parent report measures developed in the Western world are commonly used to assess children's mental health. However, previous studies have reported cultural bias in parent reports, calling into question their comparability across countries or cultural groups. The present study examines the use of anchoring vignettes to assess and adjust for bias in five countries: USA, Mexico, Germany, China, and Russia. A total of N=500 parents of underage children participated in an online survey. Parents rated their child's mental health using the Pediatric Symptom Checklist (PSC). They then rated vignettes depicting internalizing and externalizing problem behaviors. We tested for the effects of country, education, and residential environment (urban/ rural) on vignette ratings using path models. Vignette ratings were used to rescale the PSC ratings. To ensure the validity of rescaling, we tested the assumptions of vignette equivalence, response consistency, construct equivalence, and measurement invariance of raw and rescaled scores via Multi Group Confirmatory Factor Analysis. Cross-national comparisons of vignette scores revealed differences in the use of the response scale range and overall level of vignette scores (internalizing problem behaviors). Most of the assumptions for using vignettes for rescaling were met, but results for equivalence were mixed. Measurement invariance across countries improved after rescaling for externalizing problem behavior, but not for internalizing problem behavior. Rescaled scores for externalizing problem behavior revealed cross-national differences in levels of problem behavior that were masked when raw PSC scores were used. Results confirm the lack of comparability across countries in parent reports of child mental health. Anchoring vignettes appear to be a useful tool for improving comparability.

WEDNESDAY 3 JULY
Session 9.7
Topic: Identifying biases by qualitative or quantitative methods/ Translation of tests, psychological assessment instruments and survey questionnaire

126. Relationships between illegitimate tasks and employees' psychological distress: A Cross-level moderated mediating model

Yingwu Li Renmin
University of China

Drawing on Conservation of Resources Theory, this research examines the relationship between illegitimate tasks and employees' psychological distress, including the mediating roles of effort-reward imbalance and psychological contract violation, and the moderating role of perceived organizational support. A multilevel mediated effects model was constructed. Employing a time-lagged design, data were collected at three intervals using a multi-wave methodology. Participants were selected from various departments within the Guangdong Pearl River Delta, yielding 237 teams and 972 valid questionnaires. Descriptive statistics were analyzed using SPSS version 25.0, and structural equation modeling was conducted with Mplus version 8.3. The Monte Carlo method in R 4.3.1 was applied to assess the cross-level moderated effects of perceived organizational support. Findings indicate that illegitimate tasks positively predict employees' psychological distress (= 0.108, p < .05). The mediating effects of effort-reward imbalance (= 0.136) and psychological contract violation (= 0.215) are significant. Perceived organizational support mitigates the impact of illegitimate tasks on effort-reward imbalance (= -0.061) and psychological contract violation (= -0.130). This study extended the theoretical framework of Resource Conservation Theory, and redefined influence mechanism and boundary conditions between illegitimate tasks and psychological distress of employees. It also provided theoretical basis and empirical support for management practice to reduce the level of employees' psychological distress through perceived organizational support. Key Words: Illegitimate Tasks, Effort-Reward Imbalance, Psychological Contract Violation, Perceived Organizational Support, Psychological Distress

WEDNESDAY 3 JULY
Session 9.7
Topic: Identifying biases by qualitative or quantitative methods/ Translation of tests, psychological assessment instruments and survey questionnaire

135. Developing cutoff points to interpret impairment associated with depression and anxiety symptoms according to sex using the IDAS-II

Ana María de la Rosa Cáceres, Daniel Dacosta-Sánchez, Marta Narváez-Camargo, Manuel Sanchez-Garcia
University of Huelva (Spain)

The scores of the instruments that assess depression and anxiety are often interpreted using norms (percentiles). However, there is strong evidence that women score higher on depression and anxiety than men. To capture these differences, some authors have developed sex-specific norms, where the same score corresponds to a lower percentile for women vs. men. However, this approach might underestimate the severity of symptoms and the treatment needs of women, as their scores correspond to lower percentiles. An alternative to avoid this issue could be to interpret the scores in terms of their associated disability. This proposal aligns with clinical practice, in which diagnoses are established only if symptoms interfere with normal functioning. Thus, the present study aims to establish cutoffs to differentiate levels of disability associated with depression and anxiety symptoms. The participants (N = 1390) were community adults (n = 1072) and patients (n = 318) who completed the Spanish versions of the IDAS-II to assess depression and anxiety symptoms, and WHODAS 2.0 to assess disability. ROC analyses were performed to identify cutoffs in IDAS-II scores according to moderate and severe disability. According to ITC guidelines, we developed specific cutoffs for women and men to interpret the scores by considering the differences between the sex groups. AUC values were adequate (> .7) for 14 of the 19 scales for women and 17 scales for men, with the General Depression scale showing the highest ability to discriminate between moderate and severe impairments. The cutoff values for detecting moderate and severe impairments were generally higher in women. The cutoffs provided information about the disability of each assessed symptom, which could guide clinical decisions regarding treatment, drug administration, and hospitalization. Project PID2020-116187RB-I00 funded by MCIN/AEI /10.13039/501100011033, Ministerio de Ciencia e Innovación (Spain).

WEDNESDAY 3 JULY
Session 9.7
Topic: Identifying biases by qualitative or quantitative methods/ Translation of tests, psychological assessment instruments and survey questionnaire

187. Validation of depression anxiety stress scale (DASS 42) and the brief religious coping inventory (RCOPE) among Ghanaian population: Application of classical measurement and item response theories

Regina Mawusi Nugba, Enoch Tsey
University of Cape Coast

This study will assess the psychometric properties of the Depression Anxiety Stress Scale (DASS 42) and the Brief Religious Coping (Brief RCOPE) Scale among Ghanaian population through the lenses of Classical Test Theory (CTT) and the Item Response Theory (IRT). The validation study design with a quantitative approach will be adopted. A multi-stage sampling technique will be used to sample 1,536 regular undergraduate students from four select public universities in Ghana. Questionnaires will be used in data collection. Data collected will be analysed using exploratory and confirmatory factor analyses, Pearson correlation analysis, and the graded polytomous item response theory. It is anticipated that both scales will exhibit good psychometric properties among the Ghanaian population. Clinicians, psychologists and researchers will be encouraged to use the scales to enhance the diagnoses of depression, anxiety and stress among the Ghanaian population.

WEDNESDAY 3 JULY
Session 9.8
Topic: Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

94. Investigating the relative predictive validity of behaviour, personality and values upon work-related outcomes

Jasmin Kalar
Thomas International

Psychological assessments are utilised globally for selection and developmental purposes within organisations. Research has repeatedly shown 'Conscientiousness' to be one of the best predictors of success in the workplace. However, while this could be the case in conventional jobs, it may be a liability in jobs that require creativity and innovation. Some research has revealed that values can be predictive of work-related outcomes, particularly: 'achievement' and 'self-direction'. However, the research on the predictive validity of values upon work- success metrics appears to be limited, despite their applicability across cultures. Furthermore, the incremental validity of values over personality and behaviour is also underexplored. Within this research, 257 participants across six countries completed a behaviour, personality and values questionnaire, as well as work- related questions. It was found that behaviour, personality and values have predictive validity towards many work- related outcomes, although values were shown to have more predictive validity explaining some of the work-related outcomes than behaviour or personality. For example, 'intention to quit' variables appeared to be mostly explained by the value 'Family Security'. In other words, those individuals that rated 'Family Security' as lower in significance to them, were more likely to be looking for a new job. Other variables such as: work engagement, job satisfaction and organisational commitment are also explored, amongst others. It is important to understand what work-related outcomes that values can predict and whether they have more predictive validity than other measures. This research can thereby have significant implications for organisations looking to utilise assessments within their recruitment processes, as the costs of a bad hire can be detrimental. The research expands upon the current literature, suggesting that there could also be an emphasis placed upon values- based measures.

WEDNESDAY 3 JULY
Session 9.8
Topic: Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

119. Linking Personality to PsyCap: Validation of a PsyCap-based Personality Assessment using CPAI-2

Clara Y.W. To
The Hong Kong Psychological Society
Anita W.Y. Lee
The Chinese University of Hong Kong
Debbie M.W. Pong, Jack M.F. Wong, Darius K.S. Chan
The Hong Kong Psychological Society
Fanny M. Cheung
The Chinese University of Hong Kong

In a VUCA world, positivity is crucial in helping individuals adapt effectively, think creatively, and navigate challenges amidst constant changes. Positive predispositions are essential in reinforcing the impact of positive events and buffering the negative effects of adverse events. Character-building thus should be a fundamental aspect of developing psychological capital (PsyCap). Yet little research has investigated the role of personality in PsyCap. The present research aims to investigate the relationship between personality traits and PsyCap and how they impact an individual's life and academic satisfaction. Two series of studies have been carried out, with an additional goal of validating a PsyCap-based personality assessment tool. In the first study, 237 undergraduate students completed two questionnaires, namely, the Psychological Capital Questionnaire (PCQ-24) (Luthans, et al., 2007) and the Cross-Cultural (Chinese) Personality Assessment Inventory (CPAI-2) (Cheung et al., 2001). Regression analysis showed that personality traits potentially served both enhancing and protective functions on an individual's PsyCap. For instance, higher leadership tended to enhance one's PsyCap of hope, while lower inferiority protected and maintained one's self-efficacy. Path model analyses further revealed that personality traits differentially predicted the four specific components of PsyCap. The second study, which is still in progress, aims to explore whether PsyCap mediates the impact of personality on one's life and academic satisfaction. Our findings are expected to shed light on the relationship between personality and PsyCap, highlighting personality's enhancing and protective functions on PsyCap. Also, creating and validating a personality assessment tool based on PsyCap could provide a customized instrument for enhancing self-awareness of one's personality strengths and identifying areas for tailored interventions in developing each of the PsyCap components.

WEDNESDAY 3 JULY
Session 9.8
Topic: Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

253. Development and Validation of Chinese Pictorial Big Six Personality Inventory for Children (CPBSI-C)

Weiqi Mu
CAS Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China; Department of Psychology, University of Chinese Academy of Sciences, Beijing, China
Hui Cao
Capital Institute for Basic Education, Beijing Institute of Education, Beijing, China
Fugui Li
CAS Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China; Department of Psychology, University of Chinese Academy of Sciences, Beijing, China
Long Zhang
Faculty of Humanities and Social Sciences, City University of Macau, Macau, China; Student Affairs Office, Yunnan University of Business Management, Kunming, Yunan, China
Xue Li, Siying Li, Mingjie Zhou, Jianxin Zhang

Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China; Department of Psychology, University of Chinese Academy of Sciences, Beijing, ChinaThe Big Six personality structure, in addition to the Big Five personality traits, contains an indigenous trait, Interpersonal Relatedness, and is superior in explaining Chinese personality. This study aimed to develop and examine the psychometric properties of the Chinese Pictorial Big Six Personality Inventory for Children (CPBSI-C) and to examine the Big Six personality structure of Chinese children and adolescents. The inventory focused on facets under the higher-order dimensions. Each item was illustrated with pictures to make it easier for younger children to understand and raise their interest. Participants were 8,469 primary and junior high school students aged 6-17 years, and additional 177 children aged 7-12 years were recruited for a retest reliability test. The results showed that CPBSI-C had good psychometric properties, including Cronbach's α coefficient, retest reliability, inter-item correlation, structural validity, self-other agreement, external validity, and measurement invariance across gender, age groups, and the versions with and without pictures. The results also showed that the six-factor personality structure of Chinese children and adolescents demonstrated similar patterns to adults and had better goodness-of-fit indices than the five-factor structure. Using CPBSI-C, even children as young as 6 years old can self-report their Big Six personality traits, providing researchers with a consistent, more convenient, and more effective tool to study the developmental trajectory of the children's personality.

WEDNESDAY 3 JULY
Session 9.8
Topic: Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

278. Presentation and validation of the self-capacity scale, adaptation in cross-cultural situations

Nicolas Drouin, Yuanfei Huang, Anne-Marie Costalat-Founeau
France

The objective of this communication is to present the self-capacity scale based on a model, the originality of which lies in considering action and its effects as regulators of identity dynamics (Costalat-Founeau, 1999; 2021). The scale measures subjective and normative capacities, reflecting an individual's self-perceived capabilities ('I feel capable of'), and others' perceptions of their capabilities ('they find me capable of'). We tested this scale on a sample of 323 French students. An initial version of the scale included 36 items divided into 18 dyads, with each capacity being presented in its subjective and normative dimensions. We tested its internal structure through a principal component analysis and its convergent validity with DES (Differential Emotions Scale) in its French version, Orientation to life questionnaire (OLQ 13), as well as WBMMS (Well-Being Manifestations Measure Scale). Thus, we obtained a 16-item version, of which we present the psychometric characteristics. Parallely, we also adapted the capacity scale for 216 Chinese college students in France for its cross-cultural application in the fields of acculturation and well-being. The results showed that self-capacity, both subjective and normative, is positively correlated with life satisfaction and negatively correlated with potential psychological problems and perceived stress. In other words, the more capable the student feels during their studies in France, the better their mental health will be. This scale will also be applied soon in other countries for psychometric purposes. In terms of applications, this scale could facilitate guidance and support in the construction of professional projects.

WEDNESDAY 3 JULY
Session 9.8
Topic: Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

671. The Acceptance of Myths About Cyber-Sexual Violence Against Women in Spain and the United States: A Measure Invariance Study

Rocío Vizcaíno-Cuenca, Mónica Romero-Sánchez, Hugo Carretero-Dios
University of Granada

Cyber-sexual violence is a form of sexual aggression that experience the 85% of women worldwide. Research has highlighted that the myths or attitudes may play an important role to explain this form of violence. These myths have been defined as beliefs that serves to justify, minimize and deny the cyber-sexual violence against women. Specifically, the acceptance of myths about cyber-sexual violence can be assessed by the AMCYS scale, which is a unidimensional 10-item self-report scale. We studied the measurement invariance across gender and country with participants from Spain (472 men and 484 women) and the United States (502 men and 480). All participants were users of social networks and aged between 18 and 69 years old. Measurement invariance was tested with Multi-Group Confirmatory Factor Analysis (MCFA). Results showed that AMCYS was invariant at the configural, metric and scalar levels across gender and at the configural and metric levels across countries. These findings could be useful for applied researchers who want to assess myths about cyber-sexual violence, and for researchers interested in studying cultural or gender differences on these attitudes.

WEDNESDAY 3 JULY
Session 9.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

51. Item Level Exploration of Applicant Faking Behaviour on Personality Measures: The importance of Cognitive Ability and Demographic Variables

Mollie Tatlow
Thomas International/ UK

The impact of social desirability on the validity of psychological assessment in selection processes is an issue which has received significant attention. However, the importance of motivation combined with cognitive ability and demographic variables on faking behaviour at an item level has received less of a focus. When provided with the right motivational context, individuals will attempt to fake, however ability to fake and moreover, what is perceived as desirable, may be predicted by cognitive ability and demographic variables. Understanding the interaction between these variables is significant when considering fairness in testing and potential consequences to the rank ordering of candidates. Furthermore, in a globalised world, it is increasingly likely that candidates from different demographic backgrounds will be compared. 2,102 job applicants and 264 non-applicants completed the same personality and cognitive ability measure. Personality scores were explored at item level due to hypothesised differences in perceived item desirability between motivational context and demographic groups. In line with previous research, applicant scores were Inflated across most personality traits in the expected direction. Cognitive ability and demographic variables predicted how applicants and non-applicants responded on certain items within traits. Interesting relationships between cognitive ability and item scores between groups were discovered. The results indicate that the relationship between cognitive ability, demographic variables, ability to fake and actual faking behaviour may not be so clear-cut. We expand upon previous research by exploring the impact of cognitive ability and demographic variables on personality scores at an item level in genuine high stakes testing. The results of the study are discussed in terms of the potential implications of perceived item desirability on fairness in testing across demographic groups in high stakes settings.

WEDNESDAY 3 JULY
Session 9.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

179. The predictive value of the Five Factor Model across the lifespan

Nicola Taylor
University of Johannesburg
Pakeezah Rajab
JVR Psychometrics

It has been established that the Five Factor model of personality adds incremental validity to models predicting both academic and employee performance (Mammadov, 2021; Sackett et al, 2023; Wang et al, 2023). In recent years, there have been studies demonstrating how the Five Factor model can be simplified into two meta-traits (DeYoung, 2014; Strus & Cieciuch, 2019). Similarly, the facets underlying the Five Factor model can also be condensed into 10 aspects (van Lill & Taylor, 2021). This paper investigates how these various configurations of personality influence academic and workplace performance. The two samples consisted of 187 Grade 9 learners and 89 mid-level employees who completed the Basic Traits Inventory, a South African-developed questionnaire measuring the Five Factors of personality. Workplace performance was assessed using a validated performance measure, namely the Individual Workplace Performance Review (Van Lill & Taylor, 2022). Hierarchical regressions were conducted at meta-trait, trait, aspect, and facet levels. Neither of the meta-traits were significant predictors of performance in either sample, but Neuroticism emerged as a significant predictor factor, along with Anxiety, Dutifulness, and Ideas as significant predictor facets, of overall employee performance. With regard to academic performance, Conscientiousness was a significant predictor factor, Industriousness was a significant predictor aspect, and Order and Tendermindedness were significant predictor facets. This study concludes that a more nuanced approach to personality adds more value in predicting performance outcomes in both learner and employee samples.

WEDNESDAY 3 JULY
Session 9.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

215. Heterogeneity of scale items is biasing estimates of Omega consistency

Karl Schweizer, Andreas Gold, Dorothea Krampen
Goethe University Frankfurt

Consistency of the items of a scale contributes to the psychometric quality of this scale. Measures of consistency are expected to indicate larger degrees of consistency for scales composed of homogeneous items as compared to heterogeneous items. Accordingly, the reported research addressed the question whether homogeneous relationships among items positively influenced consistency measured by McDonald's Omega, that is, led to the larger estimate. Consistency estimates obtained from covariance matrices showing homogeneous relationships among items were compared with estimates achieved in investigating covariance matrices showing heterogeneous relationships. The same mean covariances characterized the two types of matrices. The investigation was conducted using ideal (= error-free) matrices and matrices computed from simulated data (200 500 x q matrices of 6 different conditions). Contrary to expectations, the general result was that heterogeneity of the relationships among items yielded the larger consistency estimates. The effect was large when the number of items was small, and it was negligible in larger numbers of items. Furthermore, the deviation showed to depend on the degree of heterogeneity. An adjustment procedure that accounts for the effect of heterogeneity is proposed.

WEDNESDAY 3 JULY
Session 9.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

605. Are Participants Fully Engaged During a Test Process? A Sequential Argument and Comparison with Different Methods

Murat Doğan ŞAHİN, Basak ERDEM KARA
Anadolu University/Turkey

Achievement tests are widely used in education to assess student performance. Ideally, in achievement testing it is assumed that students perform to the best of their ability on test items so that test scores truly reflect the construct being measured. Sometimes, however, this assumption is violated because students may not use the maximum effort necessary to process an item accurately and provide an answer that is not consistent with their true ability level. This response behavior is referred to as 'disengaged responding'. The phenomenon of disengaged responding (dr) creates a construct-irrelevant variance factor that threatens the validity of test scores. Given the potential problems that 'dr' can cause, it has become an important and interesting research topic for researchers, and several methods have been developed to detect and correct the effects of dr. Recent research aims to propose a sequential argument based upon the argument proposed by Wise, Kingsbury & Langi (2023). In their research, they examined the change in students' performance in the first and second halves of the tests by dividing the test equally. As a new argument, we will try to propose a sequential method dividing the test into two parts not equally but sequentially. In this direction, a 30-item achievement test that will be developed by the researchers and data will be collected from approximately 300 undergraduate students. For the 30-item test, we will not divide the items into two parts of 15-15 items, but 11-19, 12-18, ... 18-12, 19-11 etc. Our aim is to specify an optimal point at which disengagement begins to occur. We will then identify disengaged responders by estimating the students' ability (maximum likelihood estimation) on the two parts of the test based on our specified point. Other methods such as Mokken scaling and IRT person fit analyses will also be conducted to identify disengaged students. The consistency of these methods with our proposed method will be examined.

WEDNESDAY 3 JULY
Session 9.9
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research/ Validity theory in testing, psychological assessment and survey research

697. A Comparative Analysis of OLS and SUR Estimates in OSCE Standard Setting Using Borderline Regression Method

Haci Bayram Yilmaz
University of Aberdeen

Standard setting in Objective Structured Clinical Examinations (OSCEs) is integral to ensuring fairness and validity in medical education assessments. The Borderline Regression Method (BRM) is a widely adopted approach for establishing pass or fail thresholds. Traditionally, the Ordinary Least Squares (OLS) method is employed in the BRM standard setting. OLS estimates the passmark independently for each station, yet it fails to explicitly consider potential correlations among errors associated with the dependent variables, assuming each station's performance is unrelated to others. In contrast, Seemingly Unrelated Regression (SUR) recognizes and accommodates the potential interdependence among multiple stations. SUR captures correlations among the residuals. This study aims to compare the outcomes of OLS and SUR methods in an OSCE exam where the BRM is applied for standard setting. A cohort of 228 fourth-year medical students participated in a 9-station exam. OLS and SUR methods will be applied to calculate passmarks and standard errors of residuals for each station. To assess the accuracy of estimated passmarks, Root Mean Square Error (RMSE) will also be calculated for each station. The study will compare passmarks, standard errors of residuals, and RMSEs under OLS and SUR methods for each station. By comparing the two estimation methods in BRM standard-setting in the context of OSCEs, this research contributes valuable insights to the ongoing discourse on assessment methodologies in medical education, emphasizing the need for a nuanced approach that considers interdependencies among multiple stations.

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

540. Application of regression-based norms on the Indonesian Boston Naming Test (I-BNT) among Balinese Samples.

Aria Immanuel, Javier Suárez-Álvarez, Lisa Keller
College of Education University of Massachusetts Amherst

The Boston-naming test is a widely used neuropsychological test to assess language functioning among healthy individuals. In 2018, the test was adapted to the Indonesian language and culture, called the Indonesian Boston Naming Test (I-BNT). Yet, the current norms of the I-BNT were created using only Javanese samples. This condition threatens the applicability of the norms to other test-takers ethnic backgrounds, such as Balinese. Previous research on I-BNT found that Balinese performed lower than other ethnicities on the test after accounting for age and education differences between these groups. The first aim of this study is to apply regression-based norms among the Balinese sample. The second aim is to evaluate the consequences of violating statistical assumptions of conducting the regression-based norming study. The application of the regression-based norms is divided into three parts that consist of checking the statistical assumptions, conducting hierarchical regression analysis, and applying the regression equation lines to correct the upcoming test taker's score. This study involved the participation of 149 healthy individuals from Bali. Seven outliers were identified based on the observation of the standardized residual values. Four out of seven outliers shared common demographic characteristics, i.e., low education (below 12 years of education), and common behavior patterns, i.e., they needed a longer time to finish the test. The results of the analyses confirmed that age and education predicted I-BNT test scores. These results differed from the normative dataset from the Javanese sample that found only education as a significant predictor of I-BNT test scores. The regression equation line using the original dataset showed a lower intercept value than the outliers-excluded dataset and consequently caused different results when adjusting the test scores. The implications and limitations of using regression-based norming are discussed

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

321. Humility and Modesty: Characteristic traits of the Mexican.

Bianca Pérez
México

The main objective of the research was to provide evidence of the predictive power of the Humility and Modesty traits on trait Emotional Intelligence. The study was carried out in 2 phases where in the first phase two scales were developed, one to measure humility and one for modesty. The second phase was applied to a sample of 338 Mexican participants, of which 81% were women and the rest were men between the ages of 18 and 64 (M=26 SD 8.01). The findings indicate that the most important contribution to the prediction of emotional intelligence corresponds to modesty. It was verified through a multiple linear regression that of the two dimensions that make up modesty, the "Recognition" subdimension significantly predicts the explanation of trait EI R2.54

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

805. Evaluating Sociocultural Influences on Student Interpretation in a Science Vocabulary Measure: The Utility of a Coding Rubric.

Jose Palma
Texas A&M University
Doris Baker, Holland Briggs
The University of Texas at Austin

Theoretical Framework Understanding the role of culture in educational measurement is crucial for testing. Researchers have argued that culture is part of the validity discourse as it influences how students approach and solve items (Basterra et al., 2011). The Ecological Systems Theory (Bronfenbrenner, 2005) and the Sociocultural Theory of Cognitive Development (Vygotsky, 1978) support this view. We used these theories to create a coding rubric to evaluate how sociocultural contexts impact students' thinking when answering items about the use of science words. Objective We assess the coding rubric's utility in capturing the sociocultural contexts influencing students' interpretations of science words. The research questions are: what is the intercoder reliability when scoring responses? To what extent are sociocultural influences present in student responses? What are the sociocultural contexts in which items are interpreted based on student characteristics, content area, and item difficulty? Method Participants included 476 2nd and 3rd-grade students from 27 schools in the U.S. We used the research-developed Depth of Vocabulary Knowledge (Baker et al., 2021), and selected 20 items from the sentence production subtest. The analysis involved data from the MELVA-S project funded by the Institute of Education Studies (#R305A170455, Baker, 2020). Four researchers independently coded responses. Fleiss' Kappa was used to assess coding reliability. Descriptive and correlational analyses will present additional findings. Results & Implications Complete results and implications will be detailed in the final paper. Adapting the coding rubric from Palma et al. (2023), we identified 15 factors that range from environmental, social, and cultural influences. Initial results indicate over 90% agreement among coders. Analyzing student responses from a sociocultural perspective can enhance assessment development, instruction, and potentially higher academic success for students.

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

579. Validation of the ICF Core Set for schizophrenia from the perspective of relatives: An international study.

Karina Campoverde
Clínica-Llúria, Comunitat Terapéutica -Serveis Salut Mental (CPB-SSM), Barcelona, Spain
A.Chai Chuen
Department of Social Psychology and Quantitative Psychology, University of Barcelona, Barcelona, Spain
Juana Gómez-Benito
Group on Measurement Invariance and Analysis of Change (GEIMAC), Institute of Neurosciences, University of Barcelona, Barcelona, Spain
Maite Barrios, Georgina Guilera
Department of Social Psychology and Quantitative Psychology, University of Barcelona, Barcelona, Spain
Emilio Rojo
International University of Catalonia, Barcelona, Spain

Schizophrenia is a mental disorder that confers morbidity, disability, and a high burden of care on individuals, families, and society due to the difficulties in functioning. Evaluating functioning requires a biopsychosocial approach. The International Classification of Functioning, Disability, and Health (ICF), developed by the World Health Organization (WHO), allows describing the functioning of people with a specific health condition. This universally accepted model encompasses over 1,400 categories, which are not all specific to schizophrenia. In collaboration with the ICF Research Branch, our research team developed the Comprehensive and Brief Core Sets (CS) for schizophrenia. The ICF-CS for schizophrenia can serve as an assessment while reporting on the functioning and health in different clinical settings, countries, and cultures. The aim of the study is to describe the validation process of the ICF-CS for schizophrenia through the perspective of patients' relatives across different WHO regions. Insights from patients' relatives were collected through interviews conducted across diverse WHO regions. The sample size was established based on data saturation. Semi-structured interviews were administered, recorded, and transcribed. Following the constructivist paradigm, a qualitative content analysis was carried out, employing a deductive approach. The identified concepts were systematically linked to the predefined categories within the ICF. Twenty-seven family members from three WHO regions participated in the interviews. From the qualitative analysis, common underlying elements are obtained in how schizophrenia affects functioning, being comparable from a cross-cultural point of view. The ICF proves to be an appropriate model for obtaining information on functioning of individuals diagnosed with schizophrenia. These results largely validate the ICF categories included in the ICF-CS for schizophrenia from the perspective of the patients' relatives.

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

451. Cross-Cultural Assessment: Psychometric Requirements for a Cross-Cultural Norm of the TOP.

Jan-Philipp Freudenstein, Kilian Hasselhorn
Hogrefe Publishing Group

To ensure a meaningful cross-cultural assessment of psychological constructs researchers first analyze the degree of measurement invariance (MI) across countries. MI implies that the measurement model of a latent construct is the same across different groups, which is a prerequisite for the calculation of cross-cultural norms. The most common approach to testing for MI is multigroup confirmatory factor analysis (MGCFA). However, the degree of MI required to compare latent means across groups using MGCFA in large studies with many items, factors, and groups is rarely achieved. Alignment, an alternative approach to testing MI, introduced by Asparouhov and Muthén (2014), provides a comparison of factor means and variances across groups while allowing for approximate MI. Alignment is largely automated and allows researchers to make more sophisticated decisions about MI. With data from 12 different countries that have adapted the Dark Triad of Personality at Work (TOP; Schwarzinger & Schuler, 2016), we analyzed MI across countries using the alignment approach. Our results support approximate MI across countries and that the conditions for calculating cross-cultural norms are met. We discuss potential benefits and drawbacks of the alignment approach based on our results and provide recommendations for future researchers interested in calculating cross-cultural norms.

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

777. Herramienta docente para el desarrollo de exámenes tipo test: Construcción y validación de la Escala de Calidad de Exámenes Tipo Test.

Paula Muñoz Teno
Universidad Autónoma de Madrid

Asegurar que el sistema de evaluación universitario es de calidad es una responsabilidad ética de las universidades. Sin embargo, el proceso de construcción de pruebas válidas y fiables no está regulado institucionalmente. Existen referencias consistentes sobre las directrices que favorecen la construcción de escalas con evidencias adecuadas de validez y fiabilidad. Por tanto, la presente investigación plantea dos objetivos. En primer lugar, construir una herramienta, denominada Escala de Calidad de Exámenes Tipo Test (ECET), que permita a los docentes universitarios construir exámenes, evaluarlos y mejorar su calidad. En segundo lugar, evaluar la validez y fiabilidad de la escala. Para abordar el primer objetivo, se empleará el método inductivo y deductivo para explorar, evaluar y seleccionar la evidencia empírica sobre las variables que modulan las propiedades psicométricas de los test. Las consideraciones se agruparán en taxonomías en función de su contenido para desarrollar una escala que evalúe la calidad de un examen. Posteriormente, se realizará un pretest para desarrollar la escala final. Respecto al segundo objetivo, se realizará un estudio piloto en el que se contrastará cómo se relaciona el índice de calidad obtenido mediante la escala ECET de tres exámenes tipo test, con los indicadores de fiabilidad (alfa de Cronbach entre otros) y validez de los exámenes (análisis factorial entre otros). Los exámenes serán de asignaturas del grado de Psicología de la Universidad Autónoma de Madrid. Se espera construir una escala de calidad precisa y fiable de manera que exista una relación directamente proporcional entre el índice de calidad de ECET y los indicadores de fiabilidad y validez de los test. Generar esta escala permitirá dotar a los docentes universitarios de una herramienta eficiente, sistematizada y válida para construir exámenes de calidad, respondiendo a su responsabilidad como agentes formativos.

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

650. Development and psychometric analysis of a new scale to measure adolescents' social and public commitment for environment.

Sofia Santisi, Caterina Primi
NEUROFARBA Department, University of Florence, Italy
Angelo Panno, Luciano Romano
Experimental and Applied Psychology Laboratory, Department of Human Sciences, European University of Rome, Italy
Maria Anna Donati
NEUROFARBA Department, University of Florence, Italy

To mitigate the effects of climate change, it is urgent humans globally adopt actions that benefit the natural environment. Adolescents can have an important role in this process, as their commitment in social and public pro-environmental behaviors are fundamental in raising awareness among adults and modeling children. Although measurement of youth's environmental commitment is strategic, instruments to that aim, specifically for adolescents, are limited, and most of them focus on a unidimensional set of private behaviors. Our general goal was to provide an instrument that can be used for epidemiologic surveys and to conduct cross-cultural comparisons. This contribution regards the first step of this process, i.e., to develop a brief and multidimensional scale to assess social and public environmental engagement in adolescents, and to analyze its psychometric properties. After a focus group with adolescents to verify the items adequacy, nine items with a Likert response scale (from 1, strongly disagree to 5, strongly agree) were administered to 1826 Italian adolescents (58.1% males; Mage = 16.40; SD = 1.25). Evidence for a bifactor model with a general factor and two specific dimensions – Social engagement in promoting pro-environmental behaviors and Environmental activism – was provided. For this reason, the scale was entitled Scale for Social and Public Environmental Commitment – For Adolescents (SSPEC-A). Measurement invariance for sex and age was analyzed. Both the general and the specific factors had a good internal consistency. The total score at the scale was negatively correlated with materialism and positively related to climate change worry. Overall, this study provided a new, multidimensional, reliable, and valid brief tool for measuring social public environmental engagement in adolescents. Studies about its cross-cultural invariance are needed to analyze its potentiality in being used in reports and comparisons involving different countries.

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity and fairness in cross-cultural testing, psychological assessment and survey research

670. Harmonising the measurement of quality of life in SHARE across European regions.

Zaira Torres, Irene Fernández, Adrián García-Mollá, Amparo Oliver, José M. Tomás
Department of Methodology for the Behavioral Sciences, Faculty of Psychology University of Valencia, Spain

Background: The CASP-12 scale stands as one of the most common internationally used measures for assessing quality of life. Despite numerous validation studies, the CASP-12 factor structure remains unclear and there is no evidence supporting cross-country invariance. Objective: The aims of this study are to test the factor structure of the CASP-12 and assess its measurement invariance across four European country regions. Methodology: The sample comprised 45797 adults from the 8th wave of the Survey of Health, Ageing and Retirement in Europe (SHARE) representing four European country regions (North: Denmark, Estonia, Finland, Latvia, Lithuania, Sweden; West: Austria, Belgium, France, Germany, Luxemburg, Netherlands, Switzerland; South: Croatia, Greece, Italy, Malta, Slovenia, Spain, Cyprus; East: Bulgaria, Czech Republic, Hungary, Poland, Romania, Slovakia). Participants had an average age of 70.21 years (SD = 9.48) and 57.5% were female. A series of competitive confirmatory factor models (unidimensional, four correlated factors and second-order factor) were estimated, and a standard measurement invariance routine across European regions was used after the best fitting model was established. Results: The best-fitting model was the four correlated factors model: 2(48) = 13494.62, p < .001, CFI = 0.965, SRMR = .040, RMSEA = .078 [90% CI: .077 - .079]. This model supported configural, metric, and scalar invariance across the four European regions. Implications: Our findings suggest that the use of the four factors that form the CASP-12 (Control, Autonomy, Self-realization, and Pleasure) is appropriate, and proper comparisons can be established among different European regions.

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity theory in testing, psychological assessment and survey research

373. RPS-MAP (Route Planning Strategies in a Map) TEST: Sensitive assessment tool to identify cognitive-functional impairment related to route planning in Stroke drivers.

Candida Castro1/1, Lucía Laffarga1/2, Ana Clara Szot1/3
CIMCYC (Mind, Brain and Behaviour Research Centre), Faculty of Psychology, University of Granada, Spain
María Rodríguez-Bailón2/4
Universidad de Málaga
Daniel A. Salazar-Frías3/5, Pablo Doncel3/6
CIMCYC (Mind, Brain and Behaviour Research Centre), Faculty of Psychology, University of Granada, Spain

RPS-MAP performance-based test assess the strategic planning that takes place before driving on the road. A sample of experienced drivers [41 with a Stroke (22 Left, 15 Right side, 2 Bilateral and 2 unknown side); and 38 Healthy] was recruited. 59 men (74.68%) and 20 women (25,32%). Mean age 53.06 years old (SD=11.99). A significant negative correlation was found between RPS-MAP Total Performance & Age (r= -.24). The RPS-MAP test included 2 parts: 11 questions of planning Strategies, and 32 items of a Total Performance. 28 items of RPS-MAP Total Performance test had a discrimination index higher than .20 (4 items: 2, 3, 5 and 14, were removed). Then, Cronbach's α (.884) was found to be good. These correlations were found significant a.) Between Total Performance & Planning Time (r=.30); b.) Between Strategies & Total Time (r=.571) and c.) Between Strategies & Planning Time (r=.47). RPS-MAP test showed discriminant validity: A significant difference was found in RPS-MAP Total Performance between Healthy (65.83) & Stroke (51.79) drivers. This difference was found between Left side Stroke & Healthy drivers. RPS-MAP test showed convergent validity- These significant correlations were found significant: 1-.) Between RPS-MAP Total Performance & WCPA (Weekly Calendar Planning Activity, Salazar-Frías et al. 2023) Total Strategy Use (r=0.42), & WCPA Self-monitoring Strategies (r=0.44), & WCPA Rules Followed out (r=0.29), & WPCA Appointments Scheduled (r=0.45). Between RPS-MAP Strategies & WCPA Total Strategies (r=0.49), & WCPA Self-monitoring Strategies (r=0.33). 2-.) Between RPS-MAP Total Performance & UFOV2 (UFOV (Useful Field Of View, Ball et al., 1993) (r=.45), & UFOV3 (r=-0.55). 3-.) Between RPS-MAP Total Performance & PASAT (Paced Auditory Serial Addition Test, Gronwall, 1977) Interference Confusion Errors (r=-0.33), & PASAT Other Errors (r=-0.44), & PASAT Hits Total (r=0.36) & PASAT Errors Total (r=-0.42). Between RPS-MAP Strategies & PASAT Other Errors (r=-0.28).

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity theory in testing, psychological assessment and survey research

471. Measurement invariance of the WISC-V in Chile: A contribution to fairness in psychological assessment.

Marcela Rodríguez-Cancino, Andrés Concha-Salgado
Universidad de La Frontera

One of the most significant current challenges in psychological assessment is that the measurements made with tests should be fair and culturally relevant, given their use's impact on people's life trajectories. In this regard, the international guidelines of the AERA, APA & NCME (2018), as well as the guidelines of the International Test Commission [ITC], coincide in highlighting the importance of psychological tests having sufficient evidence on the accuracy or consistency of their scores (Reliability) on the degree to which theory and empirical evidence support the interpretations of their results (Validity) or on their ability to generate measures free of bias (Fairness). One of the psychometric strategies that can contribute to verifying the impartiality of a test is Measurement Invariance. Through this type of analysis, it is possible to elucidate whether the test scores reflect actual variations in the construct measured and are not associated with the characteristics of the group to which a test taker belongs or with biases of the instrument. This presentation will exhibit the results of a line of research in Chile that has explored whether the constructs measured in the Wechsler Intelligence Scale for Children (WISC-V) are equivalent according to sex, origin, and age group in a sample of 740 schoolchildren between 6 and 16 years of age, from the national standardization. The measurement invariance of two variants of the pentafactorial intelligence model with the ten primary subtests (hierarchical and oblique) was tested using Multigroup Confirmatory Factor Analysis. The results show complete invariance according to sex but incomplete invariance according to origin and age group, with mismatches in verbal and fluid reasoning subtests. These findings will be discussed in an attempt to contribute to good practices in the use of this instrument.

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity theory in testing, psychological assessment and survey research

143. Comparison of Threshold Identification Methods for Response Time Effort across PISA Item Types: Evaluation Based on Validity Evidence.

Militsa Ivanova, Michalis Michaelides
University of Cyprus
Hanna Eklöf
Umea University

The implementation of technology in the assessment process allows the collection of abundant process data, such as response time, which can inform whether an examinee answered an item after adequate effort. In a response-time framework that distinguishes effortful from effortless behavior, it is essential to set a threshold. This study aimed to evaluate three threshold identification methods: normative thresholds 10% and 15% (NT10 and NT15), and the change in informativeness threshold (ChInf), based on response time. Data from the computer-based PISA 2018 Reading assessment were used, including simple and complex multiple-choice, and constructed response items. The sample included 35943, 15-year-old students from Spain. Threshold identification methods were compared using five pre-defined item-level validation criteria: the ability to provide an unambiguous threshold for each item, the item accuracy rate for engaged and disengaged examinees, informativeness of effortless responses (i.e., relationship to examinee ability), and the convergent validity of the threshold identification method (i.e., the relationship with user-defined missingness). Two additional global criteria were also applied to the final examinee response time effort (RTE) scores: positive relationships were expected between RTE scores and (a) student test performance, and (b) self-reported "effort thermometer" scores. The ChInf failed to establish unambiguous thresholds for many items of each item type. The NT10 threshold identification method yielded more favorable validity results on simple and complex multiple-choice items, while the NT15 showed slightly better results on constructed-response items. Not all threshold approaches are equally valid in identifying rapid guessers and they vary depending on item type. The identification of a valid threshold identification method on various item types will allow researchers to explore changes in examinee test-taking behavior throughout the assessment.

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity theory in testing, psychological assessment and survey research

695. Validation of the Pornography Consumption Inventory in the Spanish adult population.

Oscar Cervilla Saez, Ana Álvarez-Muelas, Laura E. Muñoz-García, Pablo Mangas, Gracia M. Sánchez-Pérez, Juan Carlos Sierra
Mind, Brain and Behavior Research Center (CIMCYC), University of Granada

Introduction: The Pornography Consumption Inventory (PCI) is a scale that assesses pornography consumption. It has been validated in English and Brazilian samples and, recently, in Spanish students; however, it has not been the subject of any psychometric study in the Spanish adult population. Objective: To examine the psychometric properties of the PCI, providing validity evidence of its internal structure and relationships with other related variables and indicators of internal consistency reliability. Methods: 1,366 heterosexuals (786 men and 580 women) aged 18 to 90 years (M = 36.92, SD = 12.26) participated. A Socio-demographic questionnaire was administered to quantify the frequency of pornography consumption, the Pornography Consumption Inventory, and the Orgasm Rating Scale to assess the subjective orgasm experience of solitary masturbation in its four dimensions (affective, sensory, intimacy, and rewards). The sample was randomly divided into two groups for Exploratory Factor Analysis (EFA) (n = 400) and Confirmatory Factor Analysis (CFA) (n = 966). Correlation analysis was performed on the entire sample. Results: The EFA suggested three factors. This structure, together with the one validated in Spanish students, was tested by the CFA. The version validated in Spanish students obtained the best fit (CFI = 0.930; TLI = 0.916) and showed adequate reliability (α = .85-.87). Its three dimensions (i.e., Emotional Avoidance, Sexual Curiosity and Sexual Arousal and Pleasure Seeking) were positively related to the frequency of pornography consumption and to the dimensions of the subjective orgasm experience, with the exception of Emotional Avoidance, which was negatively related to the affective dimension. Conclusions: The three-factor version of the Pornography Consumption Inventory is confirmed in the Spanish adult population, presenting adequate validity and reliability evidence. Future analyses should extend the study to other psychometric properties.

WEDNESDAY 3 JULY
Poster Session 5
Topic: Validity theory in testing, psychological assessment and survey research

655. Psychometric Properties of the Scientific Reasoning Scale: Application to the Italian Context.

Rossella Caliciuri, Margherita Lanz
Università Cattolica del Sacro Cuore

Scientific reasoning (SR) is crucial for navigating everyday life, influencing how individuals formulate questions, gather data, evaluate information and make informed decisions. There are few tools for measuring SR, and these are tailored to specific groups and countries. This highlights the limitations of comparing results due to the unique characteristics of the tools and the need for invariant, validated instruments. The Scientific Reasoning Scale (SRS - Drummond & Fischhoff, 2017), validated in the US and Turkey (Kaygisiz et al., 2018), is a valuable multidisciplinary tool that measures an individual's ability to evaluate scientific evidence. To facilitate generalization across diverse populations, it's important to validate this scale in different cultural contexts. Our aim is to validate the SRS in the Italian context (on a sample of 600 Italian adults aged 18 and over) using SEM within the unified view of validity (Zumbo, 2005). This contemporary view treats validity as a continuum process dependent on the sample and context; as a unified concept, implying the existence of a single construct validity and different sources of evidence of validity; and implies that test users also bear the responsibility of providing evidence of measurement validity (Zumbo, 2005). Evidence will be gathered for content validity, factorial structure, generalisability, convergent and criterion validity, evidence for knowing groups, reliability. We will use some convergent measures, such as the ability to think probabilistically, open-mindedly and analytically, probability reasoning and scientific literacy. We will use attitudes towards science and scientific consensus thinking as criterion measures. Finally, we will try to understand whether the SRS is invariant with respect to gender, age, educational level, political orientation, religion, type of study. The results will contribute to the adaptation of the SRS to the Italian context and allow comparisons between

WEDNESDAY 3 JULY
Poster Session 5
Topic: Psychometric modeling

505. A B-ESEM Model for the Impact of Event Scale-Revised (IES-R): New Conceptual and Methodological Perspectives on a Popular Cross-Cultural Measure for PTSD.

Giusy Danila Valenti
University of Palermo
Palmira Faraci
University of Enna "Kore"

The Impact of Event Scale Revised (IES-R) is widely used worldwide to measure post-traumatic stress disorder (PTSD). It has been translated into several languages, with controversial findings regarding factor solution and number of items. However, the dimensionality of the IES-R has always been examined using traditional analytic strategies. Our aim was to evaluate the psychometric properties of the Italian version of the questionnaire on a sample of 231 participants (56.3% females; Mage =32.7, SD = 12.61) by testing and contrasting CFA and ESEM models. Our results showed that a B-ESEM solution was the most attractive factor structure for the 15-item version of the scale [$\chi 2$ = 52.714; df = 51; CFI = .998; TLI = .997; RMSEA = .012 (.000–.044); SRMR = .020; AIC = 8,478.705; BIC = 8,767.868; aBIC = 8,501.635], with one well-defined G-factor and three weak S-factors (Avoidance, Intrusion, Hyperousal); internal reliability coefficients were excellent (ω = .952, ωs = .952; ωh = .919; .009 < ωhs < .013). Our work supports the B-ESEM framework as an overarching perspective capable of assessing the two sources of psychometric multidimensionality of complex psychological constructs. This study also suggests that the IES-R is a robust and non-obsolete scale for assessing PTSD in current research and practice and provides a novel approach to assessing the disorder. Further studies are recommended to replicate this methodological strategy to examine the factor structure of the IES-R in different countries and languages.

WEDNESDAY 3 JULY
Poster Session 5
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

824. Creation of an International Protocol for Assessing Tests, Scales, and Questionnaires (PETEYC).

Elena Govorova, Elena de la Guía
2E, Estudios y Evaluaciones, Oviedo, Spain
Gloria García-Moreno
University of Granada, Spain
Isabel Benítez
University of Granada, Spain; Mind, Brain and Behaviour Research Center (CIMCYC), Granada, Spain

The Protocol for Assessing Tests, Scales, and Questionnaires (PETEYC) is a self-administered tool designed for professionals seeking to understand the tests' limitations and plan steps to improve the instruments' quality. The tool aims to be an open-source resource driven by collaborative efforts among practitioners and experts in psychological testing. PETEYC was created in Spanish by integrating information from previous literature and evaluations from experts in the field of educational and psychological testing. The present study aims to present the validation study conducted to gather validity evidence supporting the intended purpose and the score interpretation of the Spanish version of PETEYC, as well as presenting the English version of the tool and the information collected to identify potential limitations. First, PETEYC was administered to three assessment instruments in a pilot study. Secondly, the English version was created by following a committee approach. Finally, international experts evaluated the tool and provided information to improve its characteristics and usability. The final version of PETEYC is presented to be applied manually or digitally. PETEYC is a useful tool that will help researchers and practitioners who want to evaluate the quality of assessment instruments and will assist in making decisions about the most adequate instrument to evaluate a specific variable in a specific context.

WEDNESDAY 3 JULY
Session 10.1
Topic: Innovations in test development

268. The Validation of a Multidimensional 360-Degree Assessment Tool and Its Relationships with Supervisor Engagement, Involvement, and Burnout Tendencies

Anita Rintala-Rasmus, Mikael Nederström
Psycon / Finland

This study aimed to assess the validity of a 360 degree feedback instrument and explore its connections with leaders' commitment, involvement, and burnout tendency. The reliability & validity of the instrument were studied. Participants included leaders from top management to team leaders, with a total of 269 leaders' self-ratings & 1330 subordinate-ratings. Convergent validation explored the relationships between subordinate and leaders' self-ratings. In addition, we examined the associations between leaders' self-assessed involvement, engagement, burnout tendencies, and subordinates' 360 ratings. As an independent validation criterion, established measures of managerial work involvement (Kanungo 1982), engagement (Schaufeli et al. 2019), and burnout tendencies (Salmela-Aro et al. 2011) were used. Analysis revealed that the 360 instrument formed four independent and theoretically robust leadership dimensions, slightly deviating from Yukl's model (Yukl 2012). These dimensions measured people leadership, operational management, strategic management, and change management. Unlike in Yukl's study, external leader items were integrated within people leadership, and strategic planning emerged as a distinct factor. The internal consistency of all dimensions exceeded .90 for each scale. Furthermore, the leaders' self-reports on 360 feedback converged significantly with their subordinates' ratings. Relationships between the 360 ratings and leaders' self-reported work involvement, work engagement, and burnout index indicated that the manager's involvement is reflected in subordinates' 360 ratings, particularly on strategic management. Conversely, the burnout index or work engagement did not show significant relationships in the 360 assessments. The results imply that the investigated 360 measure is psychometrically robust, reflecting the leader's work involvement. Theoretically, the measure synthesizes previous leadership models, partly complementing them (e.g., Yukl 2012).

WEDNESDAY 3 JULY
Session 10.1
Topic: Innovations in test development

318. The identification of gifted students through a large-scale educational test: an analysis of sociodemographic characteristics

Tatiana Nakano
Pontifical Catholic University of Campinas

Special education has proven challenging in Brazil due to the difficulty in identifying the target audience. Identification is one of the most significant difficulties, which contributes to underreporting of cases. As a result of the lack of instruments and identification protocols available for this purpose, only 0.05% of students are identified. In light of this, it was proposed that a database holding the results of 100,181 students completing a standardized educational test be used to identify students who have indicators of giftedness. Each student's grade in mathematics and Portuguese was added together to give rise to one measure of academic performance. After calculating the mean and standard deviations of the total sample, participants whose cognitive performance was equal to or greater than two standard deviations above the average were identified as high cognitive performers and possible academic giftedness indicators. This strategy led to the identification of 2,149 students (2.1%). The analysis of sociodemographic data indicated a predominance of male students (51.2%), in the fifth grade (77.4%), and with an average age of 16.6 years (SD = 1.2). Students attend state schools (98.5%), in the mornings (79.9%), and have an average socioeconomic status. Based on the regression analysis, 2.1% of performance was predicted by age, 3.5% by gender, 1.3% by grade, 2.9% by the period students were enrolled in the school, and 3.3% by their socioeconomic status. Data from this study can provide information about the profile of gifted students, allowing the government to think about ways to improve the identification process in Brazil in the future.

WEDNESDAY 3 JULY
Session 10.1
Topic: Innovations in test development

392. On-screen High Stakes Assessments - Lessons learned from other jurisdictions

Jo Handford, Yasmine El Masri
Ofqual

England is examining the opportunities and risks associated with moving its pen-and-paper high-stakes assessments to the computer screen. Given the complexity and scale of the change, the Office of Qualifications and Examinations Regulation in England sought to build a deeper understanding of the benefits and challenges of such a large-scale undertaking by engaging with other countries who have progressed further on the journey. The aim was to draw relevant lessons should greater adoption of digital assessments occur in England. An initial rapid review of 50 educational jurisdictions led to a shortlist of 12 that were reviewed in more detail based on set criteria, including volume of qualifications, level of stakes of digitised assessments, availability of the documentation in English language, student performance on international assessments. Following this review, eight countries were selected for a more in-depth study. Workshops were held remotely with key individuals from relevant institutions within each country, including government organisations, examinations boards and educational institutions. The workshops probed for motivations for moving national assessments on-screen; barriers, risks and challenges in deployment; the approach of deployment, including details of implementation plans; and impacts and benefits on various stakeholders. The study suggests that countries were motivated by the potential of on-screen assessment to improve efficiency and resilience within the system, strengthen the security of assessments and improve students' assessment experience. Key challenges included a lack of consistency in information technology infrastructure and provision as well as concerns over the fairness of treatment for all students and cohorts. Ofqual is mindful of its key role in the sector in relation to enabling innovation while protecting students from any harms. This research will be key to inform regulatory approaches Ofqual adopts in the future.

WEDNESDAY 3 JULY
Session 10.1
Topic: Innovations in test development

519. **Reforming High-stakes, Low Volume Tests**

Lei Yu
U.S.

Low volume tests used for high-stakes decisions present tremendous challenges in test design, development, and validation, as established methods and procedures are typically intended for large sample sizes. The challenges are even greater for bilingual assessments where passages are presented in the target language and items in English, following the standards of the Interagency Language Roundtable (ILR) Skill Level Descriptions (SLDs) for Proficiency in Listening and Reading Comprehension. In a language system with over 150 tests in 70 languages, over half has a current volume of less than 150. The needs for test reform and refresh are urgent. Listening and Reading tests of an Asian language were selected as a case study to identify feasible solutions to an operational model. Introduced in the 1980s, the tests have 60 scored items covering the ILR scale of 1 to 3. Majority of the 100 examinees taking the tests score at 2 or above. Two-thirds of them are repeaters. The current generation tests, on the other hand, include 50 scored and 10 seeded items. In addition to reporting requirements, the ILR scale is also used to level the difficulty of passages and items. To bring the selected tests up to speed and tailor to their special features, a new test design covering the ILR range of 2 to 3 was introduced. Field test was completed, with 22 participants. Given the limited sample size, what measurement models can be used to evaluate item quality and construct the new test form? What is the optimal number of operational items to provide adequate test reliability? What standard setting methods can be used to set defensible cut scores to support high-stakes decisions (e.g., proficiency readiness, bonus pay)? How does item statistical difficulty align with the pre-assigned content difficulty? The findings of the study will provide important and rich information to operational decisions of similar tests in the broad testing field as well as the intended assessment system.

WEDNESDAY 3 JULY
Session 10.1
Topic: Innovations in test development

543. Building Culturally Sustaining Assessments to Support Adult Learners: From Co-design Studies to Assessment Development and Profile Reporting

Duy Pham, Stephen Sireci, Ketan Ketan, Eduardo Cruz, Fernando Serrano, Lian Duan
University of Massachusetts Amherst

Each year, there are millions of adult learners in the United States of America (USA) taking courses in adult education to improve their skills and uplift their quality of life and career. These learners bring with them their own cultural values, funds of knowledge, and life experience to the classroom. To help these learners, adult educators need culturally sustaining assessments that are aligned to curricula they are using and can provide actionable feedback to learners to support their learning (Suárez-Álvarez et al., 2023). To address this need, we are designing and developing a system of culturally sustaining during-instruction assessments to support the implementation of Curriculum for Adults Learning Math (CALM). We rely on Randall (2021) as the theoretical framework to inform our assessment design and development. This assessment system serves two purposes: (i) assessing whether learners show mastery of content standards and cognitive processes covered in each lesson and unit of the curriculum, and (ii) offering self-directed learning activities through assessment tasks that can provide learning feedback or be scaffolded to meet where each learner is and help them progress up to the next level of proficiency. In this presentation, we will describe our co-design studies with teachers and learners using CALM so we can collaboratively design and develop the assessment system with them. Then, we will explain our assessment design, share our assessment prototypes, and learning check reports. Finally, we will summarize user feedback from a diversity of learners and teachers using CALM who participate in our co-design studies. To the best of our knowledge, our study is among the first initiatives to build culturally sustaining assessments for adult learners in the USA. The lessons we will learn and report from this study are expected to shed more lights on how to operationalize the concept of culturally sustaining assessments and put it into practice.

WEDNESDAY 3 JULY
Session 10.3 Scholars

547. Conociendo mis Logros and AVANZO: A Comprehensive Emotional Wellbeing Assessment in El Salvador's Education System (Validity theory in testing, psychological assessment and survey research)

Fernando Mena
University of Massachusetts, Amherst

In 2020, El Salvador's Ministry of Education launched a comprehensive emotional wellbeing assessment program, targeted to 4th-12th graders. Over time, this initiative has grown to assess more than 500,000 students each year, making it the most extensive socioemotional research project to date in the country. Rooted in Bronfenbrenner's ecological system theory, the assessment examines the interplay of environmental factors on emotional well-being, combining emotional symptoms (like depression and anxiety) and socio-emotional skills from the CASEL framework. It also explores students' perception of their emotional well-being within the school, technology use, and family dynamics. The assessment uses culturally adapted Likert-type scales, refined for the Salvadoran context through a process including literature review, expert analysis, internal structure evaluation, and student interviews to ensure clear understanding, specially in the lower grades. This online assessment, part of a larger diagnostic cognitive evaluation, is voluntarily available to all students from 4th to 12th grade. The results are analyzed cross-sectionally and longitudinally, considering factors like sex, grade, school type, and location. Key findings reveal greater depression and anxiety symptoms for female students, particularly in 8th grade, an overall increase in depressive symptoms over time, and a negative correlation between socio-emotional skills and positive school perception with emotional symptoms, suggesting the former are a protective factor against the latter. Findings each year shape the next assessment and influence public policy. A "Socio-emotional Vulnerability Index" ranks schools by risk level—low, middle, or high. Resources, as psychological support and interventions, are then directed to higher-risk schools. This program's implications extend beyond assessment, paving the way for informed, data-driven strategies to support and improve student mental health in El Salvador.

WEDNESDAY 3 JULY
Session 10.3 Scholars

822. Identifying psychological factors that improve mathematics achievement in Grade 9 pupils from Gauteng (Quantitative, qualitative, and mixed validation methods)

Pakeezah Rajab
Senior Researcher at JVR Psychometrics and PhD candidate at University of Pretoria
Benny Motileng
PhD supervisor at University of Pretoria

The South African mathematics pass rate is below par when compared to international benchmarks, a trend that continues to negatively impact both tertiary education opportunities as well as the national economy. This study aimed to investigate the unique contribution of mindset, study orientations and personality traits in influencing mathematics performance, over and above the predictive value fluid intelligence adds. A sample of 187 grade nine South African pupils provided their latest mathematics marks and completed the Raven's Standard Progressive Matrices, Implicit Theories of Intelligence, Study Orientation Questionnaire in Mathematics, and Basic Traits Inventory. Logistic regressions reported that all study orientations – study attitudes, mathematics anxiety, study habits, problem-solving behaviour, and study milieu – directly predict maths marks. Additionally, the hierarchical regression models demonstrated that facets of conscientiousness, extraversion, and agreeableness moderate the influence of study orientations and predict maths performance. Overall, it is concluded that fluid intelligence, study orientations and personality add significant value in predicting grade nine pupils' maths performance.

WEDNESDAY 3 JULY
Session 10.3 Scholars

823. Five-Factor Narcissism Inventories: Psychometric Properties of the BrazilianPortuguese Versions (Translation of tests, psychological assessment instruments and survey questionnaire)

Ariela R. Lima-Costa
São Francisco University, Campinas, Brazil
Bruno Bonfá-Araujo
University of Western Ontario, London,Canada
W. Keith Campbell, Joshua D. Miller
University of Georgia, Athens, USA
Donald R. Lynam
Purdue University, Indiana, USA

Narcissism, a dimensional construct with adaptive and maladaptive manifestations, varies across cultural contexts, notably between individualistic (e.g., the US) and collectivist (e.g., Brazil) cultures. Our research explores the adaptation and psychometric properties of the Five Factor Narcissism Inventory in Brazilian Portuguese and investigates its measurement invariance with the US version. We assessed a total of 1218 participants from the general community using the FFNI (sample 1 responded to the long version, sample 2 responded to the super short version, and sample 3 responded to the super short version and the external measure) and an external measure (i.e., Pathological Narcissism Inventory). Our study involved a comprehensive analysis of internal structure, demonstrating good fit indices and reliability for the long (i.e., 148 items) and super-short FFNI versions (i.e., 15 items). Cross-cultural analysis revealed non-structural equivalence and highlighted potential cultural differences in item interpretation. Finally, our results underscore the appropriateness of FFNI usage in Brazilian samples. Despite cultural variations, narcissism scores align, emphasizing the importance of understanding cultural perception differences.

WEDNESDAY 3 JULY
Session 10.3 Scholars

220. Balancing test-taking experience and measurement efficiency in computerized adaptive testing: should easier adaptive tests be used? (Testing equivalence by psychometrics methods)

Hanif Akhtar
Doctoral School of Psychology, ELTE Eötvös Loránd University, Hungary
Balazs Klein
Faculty of Psychology, University of Muhammadiyah Malang, Indonesia
Kristof Kovacs
Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary

Although it is often claimed that computerized adaptive testing (CAT) leads to a better test-taking experience, a recent meta-analysis does not support this claim. Studies suggested that an easier CAT (ECAT), i.e., a CAT targeted at higher success rates, could positively affect examinees' psychological experiences. However, this approach is not optimal from a measurement efficiency standpoint. This study aimed to investigate the psychometric and psychological impacts of item selection in adaptive testing compared to fixed-item testing (FIT). We tested junior high school students (N = 428) in an experimental study. Participants were randomly assigned to one of three conditions, varying in test types: CAT, ECAT, and FIT. The results showed that ECAT resulted in lower anxiety, higher effort, and greater perceived performance compared to regular CAT or FIT. Additionally, the measurement precision of ECAT was superior to FIT, yet it remained inferior to that of regular CAT. Regarding testing duration, completing the ECAT required 65% less time than regular CAT or FIT. This study implies that modifying the CAT item selection algorithm to choose easier items can enhance the test-taking experience without sacrificing measurement efficiency.

WEDNESDAY 3 JULY
Session 10.3 Scholars

274. Willing and Able to Fake: A Flexible Item Response Modeling Framework for Applicant Faking Measurement (Psychometric modeling)

Siwei Peng
Jiangxi Normal University
Esther Ulitzsch
University of Oslo
Bernard Veldkamp
University of Twente
Yan Cai, Dongbo Tu, Zhichen Guo, Kai Liu, Fangbin Chen
Jiangxi Normal University

Applicant Faking (AF), the intentional distortion of responses on non-cognitive assessments to present oneself more favorably, is a critical concern in human resource management due to its impact on organizational outcomes. This study introduces a novel modeling framework for faking behavior, termed AF-IRT, which translates existing faking theory (Snell et al., 1999) into a mixture IRT model. The AF-IRT identifies faking behavior on the item-by-respondent level and decomposes the faking process into (a) participants' willingness to fake and (b) their perceptions of desirable response options. It helps explore person and item characteristics associated with higher prevalences of faking behavior and examine which response categories are seen as more desirable on a specific item. The empirical findings (within-subject design, N=586) reveal that the AF-IRT fits well with the real data and demonstrates higher reliability than existing models. A paired t-test indicates a significant difference in faking tendency (t=4.906, p

WEDNESDAY 3 JULY
Session 10.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

378. **Spanish Adaptation of the Short-Dark Tetrad (SD4)**

Jaime García-Fernández, Covadonga González-Nuevo, Álvaro Postigo, Marcelino Cuesta
University of Oviedo

Introduction: Dark personality traits are a set of psychologically aversive attributes characterized by generating individual benefits at the expense of others. Although there are different theoretical models, the Dark Triad, an aggregation of Machiavellianism, Psychopathy, and Narcissism, is the most famous theory in this field. Recently, Paulhus et al. (2021) added sadism as a new trait, constituting the Dark Tetrad, which is evaluated through the SD4. Objective: The aim of this study is to validate this scale in the general adult Spanish population. Sample: A sample of the general population was recruited through an online webpage. Methodology: The psychometric properties of the scale (item analysis, internal consistency, factorial structure) and its relationships with other dark traits, overall personality, and the light triad were examined. Results: The Spanish version of the SD4 has shown adequate psychometric properties and relationships with other variables coherent with what is stipulated in the literature. Implications: The SD4 is a valid and reliable questionnaire for assessing dark personality traits in the Spanish adults. Keywords. SD4; Dark Tetrad; Spanish; Psychometrics; Test Adaptation.

WEDNESDAY 3 JULY
Session 10.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

404. Diagnostic Adaptive Behavior Scale: Italian validation and standardization

Giulia Balboni, Alice Bacherini
University of Perugia/Italy
Pasquale Anselmi, Egidio Robusto
University of Padova/Italy

Adaptive behavior is the collection of conceptual, social, and practical skills that are learned and performed by people in everyday life (Tassé et al., 2012). Significant limitations in adaptive behavior that have emerged in the developmental period represent one of the criteria for the diagnosis of intellectual disability (ID). Developed in the USA using Item Response Theory, the Diagnostic Adaptive Behavior Scale (DABS; Tassé et al., 2017) allows the assessment of the significant limitations in conceptual, social, and practical skills in individuals 4 to 21 years old. Three DABS forms are available based on the chronological age of the assessed individual (4-8, 9-15, 16-21 years old) and are composed of 75 items each. Items are rated on a 4-point Likert scale based on how often the assessed individual performs the investigated behavior without help or reminders. This contribution presents the DABS validation and standardization process for the Italian country. This process was articulated in three main phases: adaptation to the Italian cultural context, item calibration and standardization using item response theory, and investigation of the psychometric properties. Each phase comprised several steps, including conducting diverse field tests with different committees, scale revisions, and data collection with individuals who were neurotypical or had ID. Within each chronological year (i.e., 4, 5...21 years old), recruited individuals (n = 799) were representative of the Italian population for gender and living area (North, Center, South Italy, and Islands). As the original version, the DABS Italian version resulted having excellent reliability and validity indexes, investigated with several procedures (e.g., test-retest and inter-respondent reliability, convergent/divergent validity with the Vineland-II [Sparrow et al., 2005]), as well as excellent sensitivity and specificity in detecting Italian individuals with and without ID.

WEDNESDAY 3 JULY
Session 10.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

412. Validation of the Dyslexia Screeing Test-Junior (DST-J) in an Arabic-speaking context

Mahmoud Amer Sultan
Qaboos University

It is acknowledged that Arabic is the fifth most widely spoken language in the world. Arabic incorporates both vowelized and unvowelled orthographies. Arabic words have "two independently unpronounceable bound morphemes; a root and a word-pattern," making them somewhat bi-morphemic. Consonants are expressed by symbols in unvowelled orthography, which focuses on the morphological structure of words. Empirically based screening instruments for dyslexia in Arabics-speaking contexts are very limited. The aim of the current study was to validate the empirically based Dyslexia Screening Test-Junior (DST-J) developed by Nicolson & Fawcett (2004) in an Arabic speaking context, namely Oman. A sample of 575 (289 females and 286 male) students from grades 2-5 aged 7-11 was selected (age M = 8.6, SD = 0.51). All the participants were native Arabis speakers, who were recruited from elementary schools across the country and came from low-mid socioeconomic background. The lowest scoring 20% on the LDDI (the teacher nomination survey that was used as a screener for learning disabilities) was selected to predict students' Arabic reading difficulties. The diagnostic portfolios of all the children were reviewed to make sure that these participants were still struggling with reading. It was shown that the IQ scores of all the students were within the normal range (M = 102, SD = 14). It was confirmed that none of the participant had a hearing impairment, attention deficit disorder, or a neurological/emotional disorder. From the lowest 20%, the reading, speaking, writing scales were seen as indicators of the students' struggle with reading. Content validities of the twelve scale DST-J were supported through a panel of experts. Performance of the participants on DST-J is reported, showing the at-risk quotient for each scale. Acceptable internal reliability of DST-J and test-retest reliability were found. Correlations between the 12 scales are reported. Implications for screening for dysl

WEDNESDAY 3 JULY
Session 10.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

419. Margarida Pocinho (University of Madeira), Solange Muglia Weschler (Pontifícia Universidade Católica Campinas) Thinking and Creating Styles: Assessment in Portugal

Paula Lehane
Dublin City University

The assessment of thinking and creating styles allows us to understand the creative potential of individuals to produce and excel in various areas. This evaluation provides information about the preferred ways of thinking and acting most conducive to creative production and innovation. It thus offers essential insights for talent identification, professional assessment, and vocational guidance. This study aimed to analyze the Portuguese's thinking and creative styles using the Scale of Thinking and Creative (EPC-II, Wechsler, 2023) validated with a Brazilian sample. The sample comprised 542 Portuguese adults, 63,3% female, from 18 to 75 years old (M= 33, SD=5). The EPC-II comprises 66 items, answered on a Likert scale of 6 points. The administration of the EPC-II was done online using Google Forms. This process was approximately 20 minutes, with no time limit for completion. Factor analysis using Varimax rotation yielded five thinking styles, eigenvalue>3.0, accounting for 36% of the total variance. The results indicated the presence of 5 thinking and creative styles: Cautious Reflective, Nonconforming Innovator, Logical Objective, Emotional Intuitive and Relational Divergent. Portuguese women had higher results than men (p

WEDNESDAY 3 JULY
Session 10.4
Topic: Translation of tests, psychological assessment instruments and survey questionnaire

501. Updates on the Transcultural Adaptation and Validation of the Diagnostic Adaptive Behavior Scale (DABS) for Brazil

Denise Ruschel Bandeira, Sílvia Hallberg, Adriana Schütz
Universidade Federal do Rio Grande do Sul
Euclides José de Mendonça Filho
McGill University/Canada

The Diagnostic Adaptive Behavior Scale (DABS) is a North American tool developed for the assessment of adaptive behavior (AB) of children and adolescents to assist the diagnostic process of neurodevelopmental disorders in three domains: conceptual, social, and practical skills (CD, SD, and PSD). In Brazilian adaptation, a seven-step procedure was conducted: 1) two translations of the original instrument performed by professionals in the field; 2) translations gathered and compared to get a Preliminary Translation; 3) Preliminary Translation examined by another group of professionals who verified its quality and accuracy; 4) professionals discussed the Preliminary Translation to reach the Pre-Test Translation; 5) pilot application of the Pre-Test Translation in potential users of the scale; 6) adjustments phase of the Pre-Test Translation that were identified as unclear or problematic during the pilot study; 7) administration of the final version in a large sample to determine its psychometric properties. We performed psychometric studies of item analysis, parameter estimation, validity, and reliability with a sample of 538 parents or guardians of Brazilian children and adolescents aged between 4 and 21 years old (M=10.29±4.34), 59.5% male and 66.4% from the clinical group. DABS's adapted version showed adequate psychometric properties for use in the Brazilian context. The items' factor loading ranged from .69 to .97 in CD, .61 to .96 in SD, and .48 to .95 in PSD. It presented intraclass correlation coefficients (ICC) ranging from .60 to .97, indicating a satisfactory level of agreement between evaluators and strong correlations with another AB measure already validated for Brazil. A complementary data collection was recently conducted, reaching 1,392 subjects. The next step will be the norm study. In conclusion, the adapted version will be ready for application to the Brazilian population in the following months.

WEDNESDAY 3 JULY
Session 10.5 SYMPOSIUM
Topic: Psychometric modeling

464. Forced-Choice Measurement - Challenges in Test Development

Markus Jansen
University of Wuppertal

The forced-choice method, especially the multidimensional forced-choice (MFC) format, is becoming increasingly popular for personality assessment. However, test construction with the MFC is more challenging than with traditional rating scales. There is not much evidence yet on how to effectively reduce estimation biases and improve estimation precision while simultaneously constructing tests that also take a number of practical challenges into consideration. Such challenges include different strategies for item scaling and individual trait estimation, item desirability matching, the reduction of faking, and enhancing the validity of MFC tests. Additionally, there is much variation in the utilization of models, including the framework (Factor-Analysis vs. Item Response Theory), one- vs. multidimensional tests, and model identification. The contributions in this symposium will deepen our understanding of forced-choice questionnaires and their construction using new methods of data analysis and experimental designs. Jan Killisch investigated whether the faking resistance of an MFC questionnaire can be improved by matching items to blocks using social desirability rankings instead of ratings. Christian Becker will present on a current construction of a forced-choice test for personnel selection in the Austrian military. Eunike Wetzel will discuss the construction of an MFC test based on the HEXACO-60 and show validity evidence using other personality questionnaires and criteria. Markus Thomas Jansen will present on some challenges of Thurstonian Forced-Choice modeling and introduce a new method to improve estimation accuracy by linking items between blocks. In summary, this symposium highlights challenges in Thurstonian forced-choice test construction and analysis and presents methods and empirical results to address them.

WEDNESDAY 3 JULY
Session 10.5 SYMPOSIUM
Topic: Psychometric modeling

521. Developing a forced-choice personality questionnaire for the Austrian military (Psychometric modeling)

Christian Ludwig Becker
Cupio OG – Psychologie für die Praxis, University of Vienna
Alina Bugelnig, Maria Gruber, Alexander Birner
Military Psychological Service, Austrian Federal Ministry of Defence
Susanne Frick
TU Dortmund University, Germany
Eunike Wetzel
University of Kaiserslautern-Landau, Germany

Faking in personality questionnaires is a well-known issue, especially within the context of personnel selection. The Austrian Armed Forces conduct various personnel selection procedures, all of which include the assessment of personality traits. Consequently, the current project aims to develop a standardized personality questionnaire for selecting soldiers using the forced-choice format, which has been suggested to substantially reduce faking compared to the widely used rating scale format. An existing personality questionnaire, measuring 20 personality facets derived from different job profiles and the Austrian Federal Ministry of Defence's catalog of competencies, served as the foundation for creating new items. A total of 353 positively and negatively worded items were constructed. Between April and June 2023, 3,634 conscripts (mainly 17 to 18-year-old Austrian men) filled out a preliminary version of the questionnaire using a 4-point rating scale to test its dimensionality and analyze item properties. Due to missing data and careless responding, a total of 3,041 complete cases remained. The quality of the items was evaluated mainly using CFA factor loadings and item information functions from a graded response model. The results suggest that most of the developed items are suitable for constructing a forced choice questionnaire. The next step is to create forced choice blocks using a matching procedure that accounts for social desirability. Then, data on the preliminary forced-choice questionnaire will be collected, evaluated using the Thurstonian item response model and ultimately revised. The forced choice personality questionnaire is going to be part of different selection procedures within the Austrian Armed Forces. Initially, it will be implemented in the selection process for soldiers participating in international Peacekeeping Missions.

WEDNESDAY 3 JULY
Session 10.5 SYMPOSIUM
Topic: Psychometric modeling

502. Construction and Validation of the HEXACO-MFC (Translation of tests, psychological assessment instruments and survey questionnaire)

Eunike Wetzel, Jan Killisch
University of Kaiserslautern-Landau
Susanne Frick
TU Dortmund University

The HEXACO-60 is a popular rating scale questionnaire for the assessment of the HEXACO personality traits: honesty-humility, emotionality, extraversion, agreeableness versus anger, conscientiousness, and openness to experience. Considering drawbacks of the rating scale format, we constructed a multidimensional forced-choice (MFC) version of the HEXACO-60, the HEXACO-MFC. The goal of this contribution is to describe the construction and validation of the HEXACO-MFC. We allocated the 60 items to triplets with each permutation of three out of six traits being represented once. We further took into account the keying of the items and mixed positively and negatively keyed items in each triplet. We validated the HEXACO-MFC in a sample of 940 participants who additionally filled out a number of other personality questionnaires such as the Big Five Triplets, the Big Five Inventory, and the Short Dark Triad as well as criteria including the number of Facebook friends, the number of parties they attended a month, the frequency of drinking alcohol, and their life satisfaction. Overall, convergent and discriminant validity of the HEXACO-MFC traits with other personality traits was satisfactory and criterion-related validities provided further evidence for the usefulness of the HEXACO-MFC. A comparison of the HEXACO-MFC validity coefficients with those from the original HEXACO-60 on the same variables, but an independent sample (N = 919), indicated that validity was largely comparable between the two versions. Thus, the HEXACO-MFC can be recommended for the valid assessment of the HEXACO traits in the MFC format.

WEDNESDAY 3 JULY
Session 10.5 SYMPOSIUM
Topic: Psychometric modeling

662. Matching Items by Social Desirability Rankings to Improve the Faking Resistance of Multidimensional Forced Choice Questionnaires (Translation of tests, psychological assessment instruments and survey questionnaire)

Killisch Jan
RPTU Kaiserslautern-Landau
Frick Susanne
TU Dortmund University
Brown Anna,
University of Kent
Wetzel Eunike
RPTU Kaiserslautern-Landau

The multidimensional forced choice (MFC) format requires respondents to rank items according to how well the items describe them. By matching items by their social desirability (SD), faking behavior can be reduced. Nevertheless, it was shown that faking still occurs in the MFC format. To match items to blocks, their SD is frequently assessed independently using the rating scale format. Consequently, one must assume that SD is invariant across response formats. That assumption is not necessary when the items' SD is assessed using ranking-based response formats instead. We investigated whether the faking resistance of an MFC questionnaire can be improved by matching items by joint SD rankings instead of individual SD ratings. To compare different matching methods, we re-assembled the items of a revised and extended version of the Big Five Triplets with 33 blocks. In detail, we assessed the SD of the underlying items using four methods: 1) a rating scale format, 2) a large ranking with 17 items 3) quartet rankings, and 4) a graded pairs format. Corresponding to the four methods, four different questionnaire versions with 20 blocks were created. In a separate data collection, each participant responded to one of the versions under an honest and a fake-good instruction. Using latent mean differences, we then compared the faking resistance of the questionnaire versions. We found differences in the fakability of questionnaire versions based on different matching methods. To further improve the MFC format's ability to reduce faking behavior, we propose to assess their SD using ranking-based response formats instead of rating scales.

WEDNESDAY 3 JULY
Session 10.5 SYMPOSIUM
Topic: Psychometric modeling

465. Test and item design in forced-choice modeling with Thurstonian linked blocks (Innovations in test development)

Markus Jansen, Ralf Schulze
University of Wuppertal

Thurstonian forced-choice (FC) modeling has been established as a rather new and promising methodology for scaling both items and individuals. It is most prevalently used in psychological research for the purpose of assessing constructs reflective of typical behavior, such as personality traits, and in contexts susceptible to response biases and faking. In applications of Thurstonian FC modeling, respondents typically encounter blocks containing three items and are instructed to rank the items with respect to how well the items fit as a description of themselves. None of the items are repeatedly presented in multiple blocks (i.e., blocks are unlinked). The technique is posited to substantially mitigate, if not entirely eradicate, the negative impact of response distortions on the validity of test scores. However, its implementation, especially with the unlinked block ranking format, encounters critical challenges. To address such concerns, an innovative adaptation and generalization of the block format is introduced: the Thurstonian Linked Block (TLB) design. The TLB design is applicable in both person-centric and item-centric scenarios, which makes it highly flexible. Furthermore, it also effectively facilitates a good combination of positively and negatively keyed items. Through comprehensive simulation studies, the bias in parameter estimation and the efficacy of latent trait recovery in both new linked and traditional unlinked block designs is evaluated. This is done in both factor-analytic and Item Response Theory (IRT) frameworks. The results show that conventional unlinked block designs are prone to biased results. In contrast, the TLB design exhibits mostly better performance, yielding unbiased parameter estimates, enhanced recovery of latent trait scores, and more correct model rejection rates. The implications of these results for Thurstonian FC modeling are discussed, delineating its potential in advancing psychological assessment methodologies.

WEDNESDAY 3 JULY
Session 10.6
Topic: Psychometric modeling/ Translation of tests, psychological assessment instruments and survey questionnaire

263. A Mixture IRtree Model for Aberrant Response and Missing Data

Fangbin Chen, Yan Cai, Dongbo Tu, Daxun Wang, Fen Luo, Junhuan Wei, Jiyuan Ding, Qin Wang, Zhichen Guo, Kai Liu, Xuhong Song, Pan Jiang, Siwei Peng
China

In standardized tests, examinees are likely to engage in either one or more following test behaviors: solution behavior, rapid guessing behavior, cheating behavior, nonresponse behavior, etc. Examinees do not always response all items with solution behavior due to various reasons (such as time constraint or low motivation). Aside from solution behavior, rapid guessing, cheating or nonresponse behavior can result in aberrant responses and inaccurate estimates of examinees' ability or trait, as well as item parameters, thus undermining the validity and fairness of the test. To address this issue, this paper aims to propose a mixture IRTree model to that simultaneously considers rapid guessing, cheating and nonresponse behaviors in order to model the various behaviors exhibited by examinees. The proposed model offers a notable improvement over previous studies, as it provides additional classifications for examinee behaviors at both item and examinee levels. Furthermore, it is the first model to separate and simultaneously model guessing and cheating. Two real data sets are utilized to demonstrate the reasonableness and superiority of the proposed model. Subsequently, two simulation studies based on these real data sets are conducted to validate, revealing that it provide more precise estimates of person and item parameters compared to existing models, and explored the boundary condition of model application.

WEDNESDAY 3 JULY
Session 10.6
Topic: Psychometric modeling/ Translation of tests, psychological assessment instruments and survey questionnaire

699. Dynamic Structural Equation Modeling of Daily Happiness and Stress Data

Esra Sözer Boz
Bartin University/Turkey
Derya Akbaş
Aydın Adnan Menderes University/Turkey
Nilüfer Kahraman
Gazi University/Turkey

Involving repeated assessment of individuals' current experiences in real-time and in their natural environments, ecological momentary assessment (EMA) designs have been widely used in the social sciences to collect intensive longitudinal assessment data. Dynamic Structural Equation Modeling (DSEM), combining the features of time series, structural equation, and multilevel modeling, provides a useful framework for analyzing longitudinal EMA data, given that the observations collected from the persons include a sufficient number of time points (i.e., more than 20 points). To this end, this study uses the DSEM to evaluate within- and between-person variances observed in daily happiness and stress ratings of a group of college students (n=79) to test the hypothesis that the processes would be structurally different for the two emotional experiences. To collect the data, the students were sent WhatsApp messages for 28 consecutive days, containing a link through which they would mark their daily ratings. The timing of these messages was randomized across the students to control morning, noon, and evening-hour effects. We tested a multilevel bivariate cross-lagged (1) model which included happiness and stress ratings. The results revealed that, at the within-person level, the individually standardized lagged parameters were significant, except for the effect from happiness to stress. Positive correlations were found between lagged parameters and the mean happiness and mean stress levels. The results showed that the estimated autocorrelations for stress were stronger than those for happiness and that stress levels affected the next day's happiness negatively. Our findings confirm that the dynamic processes underlying the college students' experiences of happiness and stress over time might be structurally different in that, happiness is not likely to be transferred to the next day, while stress is, i.e., unlike happiness, stress has a carry-over effect for the next day. This study was partially supported by Gazi School of Education and by TUBITAK under grant SOBAG 120K142.

WEDNESDAY 3 JULY
Session 10.6
Topic: Psychometric modeling/ Translation of tests, psychological assessment instruments and survey questionnaire

821. Beach Centre Family Quality of Life Scale: Urdu Translation, Adaptation, and Validation in the Context of Mental Illness in Pakistan

Rabia Khawar, Samavia Hussain, Mehwish Shakil
Department of Applied Psychology, Government College University Faisalabad, Pakistan
Imtiaz Ahmed Dogar
Head, Department of Psychiatry & Behavioral Sciences, Allied Hospital I, II Faisalabad, Pakistan
Ammara Butt
Associate Professor, Head, Department of Psychiatry and Behavioral Sciences, Jinnah Hospital Lahore, Pakistan
Hafiz Shafique Ahmad
Assistant Professor, Head, Department of Psychiatry and Behavioral Sciences, Nishtar Hospital Multan, Punjab, Pakistan
Hira Ahmad
Clinical Psychologist, Department of Psychiatry & Behavioral Sciences, Allied Hospital II, Faisalabad, Pakistan
Memoona Aslam
Research Assistant, Department of Applied Psychology, Government College University Faisalabad
Rizwana Amin
Associate Professor, Department of Professional Psychology Bahria University Islamabad
Ayesha Sheraz
Director, National Institute of Population Studies, Islamabad, Pakistan

Background: The Beach Centre Family Quality of Life Scale is widely used in assessing family quality of life, particularly among carers of children with disabilities (Hoffman et al., 2006). Carers of persons with mental illness also experience greater psychological distress and burden of care, resulting in decreased quality of life for individuals and their families (Phillips et al., 2023; Jeyagurunathan et al., 2017). Objectives: Family quality of life in the context of mental illness is not well studied in Pakistan, maybe due to the lack of an apt assessment tool. The current study aims to translate the Beach Centre Family Quality of Life Scale (BC-FQoL) into Urdu, adapt it, and then validate it for family caregivers of persons diagnosed with mental illness. Methodology: Approval from the original author, and the university review board was sought. We followed ITC guidelines (2017) for translating and adapting scales i.e. forward and backward translation, expert evaluation, and a pilot study with bilingual participants. Sample: After obtaining informed consent, data were collected from adult caregivers (N= 600; Mage= 38.6, SDage = 11.4; 63% women) during their visits to the psychiatric care units for their family members diagnosed with

common to severe mental illnesses in Punjab Pakistan, during March 2023 to November 2023. Results: Confirmatory Factor Analysis (CFA) was first computed on the original 25-item, five-factor scale. Finally, a 15-item, 5-factor model yielded a good fit (CFI = .96, TLI = .95, IFI = .96, NFI =.95, RMSEA =.06) as per preferred criteria (Hu & Bentler, 1999; Kline, 2015). The average variance extracted was greater than .50 for all factors with composite reliability of greater than .80 except for one factor (.78). The results are based on a single study, confined to CFA only, hence may be viewed as preliminary. Implications: The availability of the BC-FQoL Urdu version will instigate more studies and the outcomes may foster the development of psychoeducational programs for promoting family quality of life and boosting the key support systems for persons diagnosed with mental illnesses.

WEDNESDAY 3 JULY
Session 10.7
Topic: Testing equivalence by psychometrics methods

93. Measurement Invariance of the Wechsler Adult Intelligence Scale–Fourth edition across US and Spain Nationally Representative Samples

Hannah Cruickshank Campbell, Christopher J. Wilson
Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, Australia
Ana Hernandez
Research and Development. Pearson Clinical Assessment, Barcelona, Spain
Stephen C. Bowden
Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, Australia

Cross-cultural measurement invariance of a cognitive assessment is the psychometric demonstration that individuals from different cultures or countries with the same cognitive ability will achieve the same result from the same assessment. Further, establishing measurement invariance allows for the generalisability of construct validity in terms of convergent and discriminant validity cognitive construct of the assessment. Evaluating the analysis of measurement invariance of the factor structure across groups is referred to as factorial invariance. Demonstration of factorial invariance is required for statistically meaningful comparison of means across groups. This study used the Wechsler Adult Intelligence Scale–Fourth edition (WAIS-IV) standardisation samples from the United States (US) and Spain. Baseline model estimation was undertaken in the US and Spanish samples independently. A previously established four factor model showed excellent fit in both samples separately. Next, tests of factorial invariance were undertaken with the Spanish sample nested under the US sample. The results demonstrated strict factorial invariances of the WAIS-IV across the US and Spanish normative samples. These findings support the cross-cultural generalisability of the constructs as measured by the WAIS-IV. Further, the demonstration of strict factorial invariances allows for the statistically meaningful comparison of mean scores across the US and Spain.

WEDNESDAY 3 JULY
Session 10.7
Topic: Testing equivalence by psychometrics methods

117. Cross-National Generalizability of the WISC-V & CHC Broad Ability Constructs Across France, Spain, and the US

Christopher Wilson
Pearson Clinical Assessment / Australia
Stephen Bowden
The University of Melbourne / Australia
Linda Byrne,
The Cairnmillar Institute / Australia
Louis-Charles Vannier,
Pearson Clinical Assessment / France
Ana Hernandez,
Pearson Clinical Assessment / Spain
Lawrence Weiss
Test Development Consultant / USA

The Cattell-Horn-Carroll (CHC) model is based on psychometric cognitive ability research and is the most empirically supported model of cognitive ability and psychological constructs. This study is one in a series of cross-national comparisons investigating the equivalence and generalizability of psychological constructs which align with the CHC model. Previous research exploring the cross-cultural generalizability of cognitive ability measures concluded that the factor-analytic models of cognitive abilities generalize across cultures and are compatible with well-established CHC constructs. However, few studies have examined construct generalizability across non-English-speaking cultures. Specifically, the equivalence of the psychological constructs as measured by the Wechsler Intelligence Scale for Children – Fifth Edition (WISC-V) have been established across English-speaking samples, but few studies have explored the equivalence of psychological constructs across non-English speaking nationally representative samples. This study explored the equivalence of the WISC-V five-factor model across standardization samples from France, Spain, and the US. The 5-factor scoring model demonstrated excellent fit across the three samples independently. Measurement invariance was investigated, and results demonstrated strict measurement invariance across France, Spain, and the US. The results provide further support of the generalizability of CHC constructs across diverse populations and language of assessment and support the continued use and development of the CHC model as a common nomenclature and blueprint for researchers and test developers.

WEDNESDAY 3 JULY
Session 10.7
Topic: Testing equivalence by psychometrics methods

322. **Measuring well-being in school across all school ages**

Tatjana Kanonire, Daniil Talov, Maksim Kudriashov
HSE University

In our previous research, we proposed the multicomponent model of subjective well-being (SWB) in school and validated the Survey of Subjective Well-Being in School for primary school children (Kanonire et al., 2020). The model includes five components of SWB in school: satisfaction with school, affect towards school, cooperation and hostility with classmates, and subjective physical well-being, and takes into account the school context in the survey items. Due to the age dynamics of SWB, it is very important to monitor well-being at school and to be able to compare the results between children of different ages. The aim of this study was to test the multicomponent model of SWB for children of different school ages and to examine the comparability of results between children of different ages and genders. The sample consisted of 2780 students from grades 4 to 11; boys made up 47% of the sample. The Survey was modified so that it could be used not only for primary school children, but also for secondary school children. A new scale was added to measure satisfaction with the school environment (k = 8); and two versions of the survey were used - one for grades 4-6 and one for grades 7-11. The CFA, reliability and dimensionality of the scales, fit statistics and the difficulty of the statements were analyzed and the differential item functioning (DIF) analysis was carried out. The results showed that the Survey of Subjective Well-Being in School, after removing some items, has good psychometric properties and can be used to conduct comparative studies in grades 4 to 11.

WEDNESDAY 3 JULY
Session 10.7
Topic: Testing equivalence by psychometrics methods

583. Non-equivalence in PISA's ESCS index: no longer comparing apples with apples

Gavin Brown, Anran Zhao, Kane Meissel
The University of Auckland/New Zealand

PISA uses the Economic, Social, & Cultural Status (ESCS) index to quantify the level of socio-economic resources of individuals and families, and uses that index to identify economies in which student performance is greater than might be predicted by SES. One of the contributing constructs of ESCS is Home Possessions (HOMEPOS). The HOMEPOS construct was developed early in the life-cycle of PISA using indicators that relate to resources previously found to be associated with greater educational performance. However, it is possible that the substantive meaning and the predictive validity of indicators in the HOMEPOS have changed over time. For example, possession of a DVD player in the 1990s was a valid indicator of wealth but the ubiquity of streaming services no longer suggests DVD players are a sign of wealth. It is also possible that equating ESCS across jurisdictions, let alone chronologically, is no longer valid. This study investigated the construct comparability of the socio-economic status index (i.e., PISA's HOMEPOS and ESCS index) for use as a matching variable. Using data from 2012, Shanghai and New Zealand HOMEPOS items were subjected to measurement invariance testing and item response theory differential item functioning analysis. Only a few items in the HOMEPOS were statistically comparable at two pre-defined equivalency levels. These items (3 items with level 1 equivalency, and 8 items with level 2 equivalency) were treated as anchor items in developing a partially invariant socio-economic index. The validity evidence of the partially invariant index was examined in terms of its explanatory power against achievement. The results showed marginally improved variance explained by the partially invariant index compared to the original index in PISA.

WEDNESDAY 3 JULY
Session 10.7
Topic: Testing equivalence by psychometrics methods

817. Use of Computerised Adaptive Testing in the Content of Student Evaluations of Teaching at Higher Education

Ilker Kalender
Bilkent University, Faculty of Education

Student evaluation forms (SET) are higher education's most used tools to monitor instructional quality. Recently they have become online to give students more time and flexibility. No differences between two formats supported online. However, online response rates are around 50%. Students failing or forgetting to fill out forms explain low response rates. Number of items in forms is also a factor. Number of items per form may be low, but with number of courses students take, they become many. Computerised adaptive testing (CAT) is proposed as a solution. A CAT delivers a test adapted examinees dynamically, creating shorter tests with no loss of precision. Data was simulated based on the literature and characteristics from a research university in Turkey to get different class sizes (from 1, as an extreme scenario, to 100 students), the size of item pools (10 to 100 items), and number of items administered (5 to 30 items). Instructional performance was also simulated based on university's distribution as the trait to be estimated. Item parameters were estimated using Multifaceted Rating Scale Model, the novelty of this study. This model allowed to consider students' responding characteristics. Students may interpret instructors' performance differently and, intentionally or unintentionally, give lower or higher ratings. CAT simulations were conducted with several parameters. Comparisons were made using simulated instructor performance as true performance. Initial results showed reduction up to 50% in items used. More items were used to make a precise estimate of instructional performance. Also, CAT-based SET can be used graduate-level courses which have fewer students than undergraduate ones. Performance of CAT will further be compared for variables (length of SET, etc.). Results indicate CAT can be considered a solution to the response rate problem of online SET. This study is expected to provide insight into CAT in SET by creating tailored and shortened forms.

WEDNESDAY 3 JULY
Session 10.8
Topic: Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

145. The length and verbal labels do not matter: The influence of various Likert-like response formats on scales' psychometric properties

Hynek Cígler, Petra Hubatková, David Elek, Martin Tancoš
Masaryk University, Czech Republic

While the Likert scale is the most commonly used response format to measure personality traits, there is no clear consensus on how the scale's parameters moderate its performance. In two within-subject experiments, we manipulated the extremity of outer verbal labels and the presence of inner labels in a 5-point Likert-type scale (Study 1, N1 = 1044) and the scale length using 2, 6, and 10 options (Study 2, N2 = 846). We used the Height Inventory that allows for the comparison with the criterion of self-reported height and replicated the results using a typical psychological measure. In both studies, we assessed the measurement model and criterion validity. We utilized reliability analysis, path analysis, ordinal SEM, invariance analysis, and latent regressions. With more extreme outer labels and longer response scales, responses are slightly more central, impacting raw score variances (and means in skewed scales). With non-extreme labels and longer response scales, observed scores have negligibly higher reliability. Criterion validity of observed scores is only negligibly related to the presence of inner verbal labels. Reliability was higher in the all-labeled variants. We demonstrate that the measurement model can be equated across all experimental conditions, leading to an equivalent, invariant single latent trait with the same population characteristics and association with the criterion. The two-point scales resulted in lower reliability, but their criterion validity seemed unimpacted and could be advantageous in some contexts. The performance of the Likert response scale was stable across the conditions we manipulated, especially if SEM is used instead of raw score analysis. Still, we argue for verbally labeling all points on the scale and for non-extreme labels of endpoints.

WEDNESDAY 3 JULY
Session 10.8
Topic: Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

237. Examining Validity Evidence of Constructs in Applied Linguistics: A Systematic Review

Angel Arias, Mastoor Al Kaboody
Carleton University / Canada

Research instruments, such as questionnaires and tests, are pervasive in applied linguistics and related social science sub-disciplines. Notably, high-stakes language proficiency tests like TOEFL and IELTS boast robust validation and research programs for defensible score interpretation and use (Kane, 2013). While language proficiency, particularly in English, undergoes extensive validation research, other constructs, such as willingness to communicate, grit, and resilience, may lack comparable validity evidence. This discrepancy raises concerns about the accurate interpretation and practical application of data derived from these research instruments. Consequently, this emphasizes the need for continued scrutiny and validation across diverse constructs in applied linguistics. We applied the conceptual framework in the Standards for Educational and Psychological Testing (AERA et al., 2014) to evaluate the validation of research instruments in peer-reviewed journals of applied linguistics between 2010 and 2022. Employing a mixed-methods approach in line with PRISMA methodological guidelines (Page et al., 2021), we analyzed 448 studies, incorporating 71.65% close-ended and 21.65% open-ended questionnaires, with a focus on adapted instruments (70.76%). This revealed a strong preference for quantitative survey methods, with less frequent use of qualitative approaches. Evidence based on test content and the internal structure of instruments were conducted in 40.63% and 43.75% of studies, respectively. However, there was a lack of validation, encompassing response processes, relations to other variables, and consequences of testing. Our findings emphasize the necessity for context-specific validation. Relying on adapted tools prompts questions about repurposing—using research instruments in contexts they were not originally designed for. This underscores the need for validation efforts to address underrepresented validity evidence for key constructs in applied linguistics

WEDNESDAY 3 JULY
Session 10.8
Topic: Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

562. Applying Multidimensional IRT Models in Validating the Dimensionality of the Future Skills Exam

Fathima Jaffari
Senior Measurement specialist, department of Tests and Measurement, National Center for Assessment, Education and Training Evaluation Commission (ETEC), Riyadh, Saudi Arabia
Rehab AlHakmani
Emirates College for Advanced Education, United Arab Emirates

Employment skills are necessary for the successful engagement in the labor market and they are as important as the academic and technical skills that undergraduate students learn at university. There is an emerging trend that employers will begin to shift towards a broader skills-based perspective in employment selection as a substantive shift over the past few years (World Economic Forum, 2020). Therefore, the national center for assessment (NCA) has developed the future kills exam to measure the employment skills and competencies that are required by the Saudi labor market. The purpose of this study was to explore the dimensionality of the future skills exam using a multidimensional item response theory (IRT) framework. Two forms of the exam were developed where each form consisted of 50 multiple-choice questions grouped into two domains and five skills. The two forms were administered to senior students at different university. In particular, form 1 was administered to 11, 234 students while form 2 was administered to 10, 156 students. To validate the structure of the exam and explore possible multidimensionality, a hierarchical multidimensional IRT analysis was implemented, based on the domains and the skills measured by the exam. The nested models were tested for significant differences using chi-squared deviance test and the best was selected using the Akaike's information criterion (AIC; Akaike 1974). The results suggested that the two-dimension model was better in structure than both the unidimensional model and five-dimension model. Further, the results indicated that the two-dimensional model had good empirical reliability across the two forms where the two dimensions (i.e., cognitive and social-emotional) were strongly correlated. Since the exam was designed to assess the employment skills that are required by the Saudi labor market, measuring both cognitive and social-emotional domains is crucial as these are equally valued in the labor market.

WEDNESDAY 3 JULY
Session 10.8
Topic: Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

705. Generative item models as learning opportunity: a psychometric analysis of 25 item models in the mathematical domain of space & form

Philipp Sonnleitner, Pierrick Kinif, Caroline Hornung, Steve Bernard, Ann Kieffer
University of Luxembourg/ Luxembourg

Despite the disruptive impact of generative AI on item writing, there are still merits to pre-defined, template-like item models that can be used in a generative way too. These models offer transparency, as they are not a black box and can be validated by subject matter experts. They also allow for predictable characteristics in generated items and facilitate systematic exploration of the models' psychometric properties, yielding valuable insights into the targeted construct. Focusing on the mathematical domain of space and form, this study psychometrically analyzes 25 cognitive item models developed by a team of national experts. These models were used to create items for Grades 1, 3, and 5 in the Luxembourgish school system. Each model underwent administration in six experimentally varied versions to assess the impact of contextual presentation and problem characteristics identified by cognitive psychology as influential in problem-solving processes. Data from the annual school monitoring Épreuves standardisées for Grade 1 (n = 3694), Grade 3 (n = 4625), and Grade 5 (n = 3716) was analyzed in-depth by descriptive comparisons of resulting IRT parameters, and the estimation of manipulated problem characteristics' impact on item difficulty by using a generalized linear mixed model with students as random effect (GLMM, De Boeck et al., 2011). This allowed for an evaluation of the stability, predictability, and unbiased nature of the psychometric characteristics of items generated by each model, particularly concerning subgroups known to be disadvantaged in the Luxembourgish school system. The findings offer significant insights into the space and form domain in mathematics, a vital but less studied area. They reveal how the graphical nature of items in this domain is substantially influenced by their presentation, underscoring the importance of controlled, template-based item generation for ensuring versatility and fairness.

WEDNESDAY 3 JULY
Session 10.8
Topic: Validity theory in testing, psychological assessment and survey research/ Validity and fairness in cross-cultural testing, psychological assessment and survey research

808. Testing Individuals with Disabilities for International University Admissions: Learning from the United States Experience

Kurt Geisinger
Buros Center for Testing, University of Nebraska-Lincoln, USA
Agustin Barroilhet
University of Chile
Monica Silva
Catholic University of Chile.

The aim of this paper is to identify key aspects of the American experience in testing individuals with disabilities in university admissions to develop a protocol, based on the ITC Guidelines and AERA, APA, and NCME Standards potentially applicable in Chile and other Latin American countries. This process entails not only "the proper accommodations that the individual needs to take a test in a manner that reflects their knowledge and ability rather than their disability" (Geisinger, 2022), but also scoring and sharing these scores balancing the privacy rights of individuals and the needs of information of higher education institutions. Most Latin American countries do not have statutes regulating the testing of students with disabilities. Each country has taken a local approach based on the capacity of the agencies designing test and/or court judgments. For example, Colombia's Saber 11 test was recently reformed by the Colombian Institute for Educational Evaluation [Instituto Colombiano para la Evaluación de la Educación] after a ruling from the Columbian Constitutional Court. This work aims to acclimate international testing professionals and audiences in Chile to the particular challenges entailed in testing individuals under The aim of this paper is to identify key aspects of the American experience in testing individuals with disabilities in university admissions to develop a protocol, based on the ITC Guidelines and AERA, APA, and NCME Standards potentially applicable in Chile and other Latin American countries. This process entails not only "the proper accommodations that the individual needs to take a test in a manner that reflects their knowledge and ability rather than their disability" (Geisinger, 2022), but also scoring and sharing these scores balancing the privacy rights of individuals and the needs of information of higher education institutions. Most Latin American countries do not have statutes regulating the testing of students with disabilities. Each country has taken a local approach based on the capacity of the agencies designing test and/or court judgments. For example, Colombia's Saber 11 test was recently reformed by the Colombian Institute for Educational Evaluation [Instituto Colombiano para la Evaluación de la Educación] after a ruling from the Columbian Constitutional Court. This work aims to acclimate international testing professionals and audiences in Chile to the particular challenges entailed

in testing individuals under special accommodations and avoid, if possible, its judicialization. In the U.S. there is a relatively vast legislative record in this area, including historical documents that shed light on the Rehabilitation Act´s rocky beginnings. These reveal that it was the subject of intense debate and compromise on the part of Congress and the President and its aftermath. After the bill was signed and became law, cases have been brought to court, among them cases in California that deterred flagging practices by the College Board, ETS, and the Law School Admission Council (Geisinger, 2022). It now appears that no major testing agencies continue to flag test scores of students with disabilities in the United States.