# Testing International

**Volume 39, June 2018**
**Editor: April L. Zenisky**



**Testing International is a publication of the International Test Commission**

## Content



*Are you ready for the 11th Conference of the ITC?*
*Turn to Page 3 for a preview of upcoming conference*
*(July 2 to 5, 2018) in Montréal, Quebec, Canada*

# Greetings from the Editor

**April L. Zenisky**
Editor, *Testing International*
University of Massachusetts Amherst

I hope this issue of *Testing International* finds you well. I write you as the academic year is winding down at UMass Amherst, in the USA, and the weather in my corner of the world is finally becoming sunny and warm. This column is presented to you with mixed feelings, as this is my final issue as Editor of *TI*, but I know if will be in *very* good hands with the next Editor, Dr. Nicola (Nicky) Hayes. Being Editor of *TI* has been an experience that I have enjoyed very much, and I wish Nicky all the best as she takes on this newsletter and the many opportunities this work offers to interact with the ITC membership and the international testing community. By way of introduction, Nicky has contributed a book review to this issue, which you can find on Page 21.

Also going on as I have been putting together this issue are the final preparations for the 11th ITC conference in to be held in Montreal, Quebec, Canada, July 2-5. Montreal is a beautiful and vibrant city, and to add to the excitement of the conference, the internationally renowned Montreal Jazz Festival will *also* be happening then.

The conference organizers have been incredibly busy putting together the program, communicating with members and presenters, and putting the finishing touches on all the details that managing a conference of this scale requires. On Pages 3-5 of this issue they've provided some information about the truly impressive slate of workshops and keynote speakers they have assembled. It should be an excellent conference, and you can find more information on the conference website: https://www.itc-conference.com/

There are some other highlights of this issue that I wish to bring to your attention: Included here (pages 6-20) are reports from most of the 2018 ITC Scholars. The Scholars program is a critical part of the outreach of the ITC and highlights the accomplishments of early career scholars working in the areas of psychological and/or educational testing from developing and emerging economies.

Also presented here is an update on the recent activities of the International Association for the Evaluation of Educational Achievement, including information about their initiatives and many links to resources (p. 31-34). In addition, a collaborative research team from Australia has shared with the ITC community some of their research on choice with respect to open-ended questions (p. 23-31).

Avi Allalouf from NITE has provided us with some details on the steps being taken in Israel to advance psychometric related training, both with regard to traditional graduate studies and through the use of MOOCs (p. 38-41).

And finally, in recent years, the ITC has been asking individuals who have held various leadership roles in the ITC to respond to a brief series of questions about their service, which are being published on both the ITC website and here in *TI*. In this issue (pages 34-37) I am very pleased to share with you the Early Leaders Interview with my graduate advisor and now colleague at UMass, Dr. Ronald K. Hambleton, who has been a leading presence in the ITC for many, many years. I hope you enjoy reading Ron's perspective on the ITC's evolution from its earliest days to the present!

# Bienvenue à Montréal!


## The International Test Commission Conference is approaching!

Dear Colleagues,
We are delighted to invite you to register for the ITC Conference workshops, which will be held on July 2, 2018 at Centre Sheraton Montréal.

***A unique opportunity***
Our workshops will be facilitated by experts who are recognized around the world for their theoretical and practical expertise in their field. They are intended for researchers, practitioners and students who wish to receive training and share their experiences on a variety of important subjects.

In total, 11 workshops (6 lasting a whole day and 5 lasting half a day) will be offered, focusing on such topics as evaluating collaboration in problem-solving, standard setting, various psychometric methods, and testing-related ethical principles.

Here are all of our workshops (link)

**Early-bird registration fees will be maintained until May 30, 2018! Don't be late!**

- Are you already registered for the conference, and would like to attend a workshop? No problem! Attend a workshop
- Not registered for the conference yet? Right this way. You will be able to choose workshops at the same time.
- Would you like to attend a workshop without participating in the conference? You can also do that here: Attend only a workshop
- You can also book your hotel room

We look forward to welcoming you in large numbers!

---

## *Montreal 2018 Keynote Speakers*

**Alina von Davier**
Vice President of ACTNext, by ACT, Inc., Research, Development, and Business Innovation Division, Adjunct Professor at Fordham University, USA

- *The Application of Computational Psychometrics to Process Data from Performance Assessments*

**André De Champlain**
Director, Psychometrics and Assessment Services Department, Medical Council of Canada (MCC), Canada

- *Implementing Automated Item Generation in a Large-scaled Medical Licensing Exam Program - Lessons Learned.*

**Bruno Zumbo**
Professor and Distinguished University Scholar, the Paragon UBC Professor of Psychometrics and Measurement, University of British Columbia, Vancouver, Canada

- *The Reports of DIF's Death are Greatly Exaggerated; It is Like a Phoenix Rising from the Ashes*

**David Magis**
Research Associate, Fonds de la Recherche Scientifique (FNRS), Department of Education, University of Liège, Belgium

- *Adaptive Testing: Examples, Simulations, and Examples of Simulations*

**Déon de Bruin**
Professor, Industrial Psychology, University of Johannesburg, Head of the Centre for Work Performance, South Africa

- *Challenges of Psychological Testing in the Multicultural South African Context*

**Irini Moustaki**
Professor, Social Statistics, at the London School of Economics and Political Science

- *The Contributions of Women in Psychometrics-Statistics: Past and Present*

**Leslie Rutkowski**
Professor of Educational Measurement in the Centre for Educational Measurement, University of Oslo, Norway

- *Increased Heterogeneity in International Assessments and Associated Measurement Challenges*

**John Hattie**
Laureate Professor and Director of the Melbourne Education Research Institute at the University of Melbourne, Australia

- *Visible Learning and Assessment*

**Maryam Wagner**
Assistant Professor, Department of Medicine, McGill University, Montreal, Canada

- *Examining the Potential and Uses Cognitive Diagnostic Assessment in Test Development and Validation*

**Stephen G. Sireci**
Professor, College of Education; Director, Center for Educational Assessment, University of Massachusetts Amherst, USA

- *21st-Century Validation Procedures for 21st-Century Tests*

**Sara Ruto**
Head Secretariat, People's Action for Learning (PAL) network, Kenya

- *Including the Excluded through Rethinking National Assessments: The Example of Citizen Led Assessments*

## ***Montreal 2018 Workshops***

**Assessment of Collaborative Problem Solving Skills: An Overview**
*Alina Von Davier, Kristin Stoeffler*

**Ethics, Test Standards and Test Interpretation: Measurement Matters**
*Gary L. Canivez*

**Introduction to Automatic Item Generation using CAFA AIG**
*Jaehwa Choi*

**Cognitive interviewing for interpreting DIF form a mixed methods perspective**
*Jose-Luis Padilla, Isabel Benítez*

**Generalizability theory: Application and optimization**
*Lisa A. Keller, Robert J. Cook, Frank V. Padellaro*

**Applying the Standards for Educational and Psychological Testing in International Contexts**
*Linda Cook, Wayne Camara, Joan Herman, Kadriye Ercikan*

**Tools for Equating**
*Won-Chan Lee, Kyung (Chris) T. Han, Hyung Jin Kim*

**Quality Control Procedures for the scoring and rating of Tests**
*Avi Allalouf*

**Tests and Score Report Design Principles for Culturally and Linguistically Diverse Populations**
*Maria Elena Oliveri, April Zenisky*

**Applying Test Equating methods using R**
*Marie Wiberg, Jorge Gonzalez*

**Crafting adapted tests with a focus on a-priori methods**
Dragos Iliescu

## Tourism Montréal

*Fun Fact #1:*
Montréal is the second most-populous city in Canada.

*Fun Fact #2:*
The city is named after the triple-peaked hill in the heart of the city, Mount Royal.

For more info about Montreal, go to: https://www.hikebiketravel.com/30-fun-weird-interesting-facts-montreal-quebec/

## Next Generation Technology-Enhanced Assessment: Global Perspectives on Occupational and Workplace Testing


**Editors**:
John C. Scott, APT Metrics
Dave Bartram, CEB-SHL
Douglas H. Reynolds, Development Dimensions International

The use of technology for workplace and occupational testing blossomed in the early years of this century. This book offers a demonstration that the first generation of these technologies have now been implemented long enough to observe the patterns and issues that emerge when these approaches evolve through technical advancement and successive application. A new set of issues and opportunities has emerged and the next generation of these applications is now coming of age. This book reflects on the last few decades of this evolutionary process from a vantage point of global experience across a wide range of workplace applications, including employment selection, development, and occupational certification. This text provides an essential foundation for individuals involved with the assessment of human capability and potential in organizational and workplace contexts.

- Covers the next generation of technical advancements in workplace testing
- Draws from experts around the world and across practice domains
- Offers a practical perspective for real-world application

Read more at
http://www.cambridge.org/gb/academic/subjects/psychology/applied-psychology/next-generation-technology-enhanced-assessment-global-perspectives-occupational-and-workplace-testing#ZSY2UEHJhRmA6fej.99

## Competencies of the 21st Century in Brazil

**Carolina Rosa Campos**
Universidade São Francisco, Brazil

Education is going through time of fast and deep changes related to the functions of teaching and school programs. Many countries around the world are rethinking their educational systems in order to prepare the new generation to the 21st century and that's not simply a national matter, but a global challenge.

Reflecting and restructuring methods and teaching-learning proposals deal with the importance of the development of socioemotional competences at the school environment in order to stimulate the complete development of children, teenagers and young adults. Known as 21st century competences, socioemotional competences go beyond learning how to read and write, math and other sciences, intertwining the importance of learning how to collaborate, persist, organize, create, overcome and build up confidence and optimism in human relations.

In Brazil, most efforts are being put on reorganizing the Base Nacional Comum Curricular (BNCC), which defines a group of 10 general competences that must be developed alongside the curricular components throughout basic education. At the same time, the focus on students' development has also been watched in large-scale evaluative processes, via studies with PISA (Programme For International Student Assessment) and ENEM (Exame Nacional do Ensino Médio).

The mobilization in watching the competences' progress through the school year, as well as identifying their evolution has motivated significant research in the country. In Brazil, Instituto Ayrton Senna (IAS), by the Edulab21, has started to create and implement educational solutions in order to fully develop the new generations' potential, allowing us to build a better world.

Related to that, it is relevant to highlight the studies with the SENNA instrument, a test based on the Big Five model, elaborated with 17 characteristics and five domains in which the student reflects about itself (its behavior in certain situations). The answers represent an indicator of the Big Five Personality Traits (extraversion, conscientiousness, openness to new experiences, agreeableness and neuroticism) and a sixth aspect, called Locus Of Control, which reflects in what measure the individual assigns situations to attitudes taken by it, or to incidents and decisions taken by others.

The preliminary studies with the instrument show optimistic results, indicating that students with more developed socio emotional competencies are prone to have a better school performance and that it is possible to stimulate these competencies with intentional actions, by public policies. In this case, it is understood that competencies such as perseverance and curiosity complement the role of cognitive competencies for school development, but they must be stimulated from basic education until the students are inserted in citizenship and work environments.

To achieve this, the institute has been working on complementary projects so as to try to

supply some demands connected to the 21st century competences inside school context. The creation of a creativity and critical thinking formative evaluation instrument, for instance, has the focus of, in the future, provide support to teachers while elaborating pedagogical processes which promote the intentional and well-built development of these competences. In the same way, the Programa de Letramento em Computação, held by the IAS in a partnership with Universidade São Francisco (USF), aims to endorse the literacy in programming languages in service of the computational thinking development with basic education public school students, being able to reflect on the development of other hybrid skills, such as creativity.

These actions and research intend to demonstrate the importance of the comprehension of cognition, practical skills, socioemotional and physical as a whole and with the potential to influence, in a more positive and representative way, the teacher-student relation, when applied in a harmonious way. In a more complex way, they intend to emphasize that the stimuli and insertion of these allied competences may offer a strong impact for the development of cognitive skills and participate in personality strengthening.

It must be clear that building up changes in Brazil's educational system is not an easy task, but it may become more accessible when thought through a comprehensive sight and considering different variables, which can be directly associated to students' learning styles in the classroom. In this case, it's more important for each student's individuality to be respected, and that we seek, starting with the singular sight, develop strategies to make education more efficient.

## The application of psychological measurements on research purposes: a systematic review about mental health problems among adolescents in Vietnam

**Anh Nguyen**
School of Public Health and Preventive Medicine
Monash University, Australia

Psychological assessment is a systematic evaluation of the current level function of a person (Haynes, 1998). In research context, psychological assessment plays an important role in identifying and explaining behavioural, psychological, and cognitive signals of a person; which in turn can significantly affect the study's conclusions and the suggested strategies for further considerations or treatment plans (Haynes, 1998). Psychological assessment, therefore, is considered as an essential component in the evaluation of a study quality (Critical Appraisal Skills Programme, 2017).

Vietnam is a lower middle-income country in South East Asia with an area of 362,000 square km (Cao et al., 2016; The World Bank, 2011). It has a population of more than 90 million people, comprising 54 ethnic groups (General Statistics Office of Vietnam, 2016). Vietnamese society was influenced by Confucianism and feudalism in China for more than 1000 years (Vu, 1997). It had also experienced many wars to gain national

liberation until 1975. In 1986, a new policy was announced which led to significant economic and social changes. A mix of private and public sectors and the market-based economy was established resulting in a significant boost to the national economy (SAVY I report). During this time, psychology achieved an important milestone with the establishment of the Institute of Psychology in Hanoi in 1993 (M. H. Pham & Do, 2004). Since then, the discipline of psychology has made considerable progress with many scientific publications within and beyond the country. However, this relatively young field of science continues to face many challenges, one of which is the use of validated psychological measurements in research. This article establishes the available evidence of using psychological measurements in studies investigating mental health problems among adolescents living in Vietnam.

A systematic search was conducted on widely used electronic databases: Medline, Embase, Cinahl, PsycINFO, Scopus and Web of Science from 1st January 1993 to 1st May 2017. Handsearching of Vietnamese printed journals and contacting authors were used in addition to maximizing the opportunity of including all potential studies. The key- search terms were “mental health problems”, “adolescents” and “Vietnam”. Studies were included if they were published in English or Vietnamese, investigated the prevalence and/or determinants of mental health problems among adolescents aged from 10 to 19, and had been living in Vietnam.

Of the 2,749 records retrieved 20 studies were included in the review. There were 14 published in English and six in Vietnamese. Among these, thirteen used quantitative research methods, including three secondary analyses of data from the National Survey of Vietnamese Adolescents and Youth (SAVY) (Kaljee et al., 2011; L. C. Le & Blum, 2015; M. T. Le, Nguyen, Tran, & Fisher, 2012), one cross-sectional survey using structured interview with parent/guardian (Amstadter et al., 2011), one cross-sectional using a parent-report questionnaire (McKelvey, Davies, Sang, Pickering, & Hoang, 1999), and eight cross-sectional surveys using self-report questionnaires (M. T. H. Le, S. Holton, H. T. Nguyen, R. Wolfe, & J. Fisher, 2016; D. Nguyen, Dedding, Pham, Wright, & Bunders, 2013; H. T. Nguyen, Dunne, & Le, 2009; T. T. P. Pham, 2014; Thai et al., 2015; B. P. Tran, Nguyen, Truong, Hoang, & Dunne, 2013; T. N. Tran, 2015a, 2015b). One paper reported the qualitative component of a mixed-method study (D. T. Nguyen, Dedding, Pham, & Bunders, 2013). The remaining six papers reported mix-method studies using both quantitative and qualitative sections (B. D. Nguyen, 2014; D. Nguyen et al., 2013; H. T. Nguyen, 2006a; Hue Thi Nguyen & Nguyen, 2012; T. T. B. Pham, 2015; V. T. Pham, 2016; Stratton et al., 2014). In total, 24 questionnaires, scales and checklists were used to investigate mental health problems among adolescents in Vietnam, including self-reporting and parent-reporting tools.

Five of 24 assessment tools used in 20 papers were adapted from English into Vietnamese by translating and back-translating; cultural sensitivity was tested by pilot studies. The most commonly used screening tool was the Center for Epidemiological Studies Depression Scale (CES-D) for assessing symptoms of depression. The cut-off point for having depressive symptoms among Vietnamese adolescents was 16 point (D. Nguyen et al., 2013; H. T. Nguyen, 2006b; H. T. Nguyen et al., 2009; T. T. B. Pham, 2015; Thai et al., 2015). However, a 2005 study of Nguyen and his colleges used a cut off point of 21 for depressive symptoms and 25 for severe

depression symptoms. This is problematic as the cutoff point of 21 had never been used and was likely to under-report depressive symptoms in adolescence living in Vietnam. The Educational Stress Scale for Adolescents (ESSA) (D. Nguyen et al., 2013; T. T. B. Pham, 2015; Thai et al., 2015) was used in three reports and had construct and concurrent validity on 1,226 adolescents attending schools (Thai et al., 2015). The Strengths and Difficulties Questionnaire (SDQ-25) was used in two studies (Amstadter et al., 2011; Stratton et al., 2014). The validity of this questionnaire has only been established to be used among adults in Vietnam (Giang, Allebeck, Kullgren, & Nguyen, 2006; T. Tran, Harpham, & Nguyen, 2004). There was no available information about the SDQ-20 validation on Vietnamese adolescents. The Depression, Anxiety and Stress Scale – 21 (DASS-21) was used by Le in a 2016 study (T. H. M. Le, S. Holton, T. H. Nguyen, R. Wolfe, & J. R. W. Fisher, 2016); however, this scale has been validated among adults only (D. T. Tran, Tuan, & Fisher, 2013). The Anxiety Scale was reported as a validated measurement in two studies (D. Nguyen et al., 2013; Thai et al., 2015); however, the validation paper of this instrument was not accessible.

All studies used measurements that were reported to be validated in Vietnam followed standard steps in cross-culture validation. Translation and back translation by separated bilingual psychologists were completed, followed by a pilot study on target population before conducting a main survey or interview. The sample size for five papers establishing validation ranged from 22 to 1,226 participants. All had received ethics approval for conducting the studies.

The national surveys to obtain baseline data about Vietnamese youth in 2003 and 2008 used the questionnaires named SAVY I (General Statistics Office of Vietnam, 2005) and SAVY II (General Statistics Office of Vietnam, 2008). These are the only two surveys for adolescents developed in Vietnam with the collaboration between the Vietnamese Government's Ministry of Health (MoH), the General Statistics (GSO), the WHO, the United Nations Children's Fund (UNICEF), Vietnamese Ministry of Education and Training (MoET), the Vietnamese Central Youth Union (YU), and the Vietnam Women's Union (VWU) (General Statistics Office of Vietnam, 2005). However, no descriptions about pilot studies were published.

Within 20 reviewed papers, 17 of 24 questionnaires used instruments that were not validated in Vietnam, including the Child Behaviour Checklist (CBCL) (McKelvey et al., 1999) (T. T. P. Pham, 2014), the World Health Organization (WHO-5) (T. T. B. Pham, 2015; Thai et al., 2015), the Kessler Psychological Distress Scale (K-10) (T. T. B. Pham, 2015; Thai et al., 2015), the Youth Risk Behaviours Survey (YRBS) (T. H. M. Le et al., 2016), the Carl Jung's Anxiety Scale, the H.J. Eysenck Personality Questionnaire (Hue Thi Nguyen & Nguyen, 2012), the Muris' and Myers' Anxiety Scale, the Behavioural-Emotional Disorder Youth self-report (T. N. Tran, 2015b), the Parental Authority Questionnaire- PAQ, the Child's report of parental behaviour Inventory - CRPBI [16], the Spitzer's General Anxiety Disorder (GAD-7), the Phillip's Anxiety Scale, and the Rosenberg's self-esteem Scale (T. N. Tran, 2015a). These questionnaires had commenced cultural adaptation by translating the tool from English to Vietnamese; however, no further steps were reported.

Academic Motivation Scale (T. N. Tran, 2015a), Self-report Post-Trauma Stress Disorder Scale, and Difficulties Questionnaire

for Adolescents, Self-Report Domestic Violence Scale (B. D. Nguyen, 2014) were mentioned in two studies without references and authentications.

This review reveals that most psychological assessment tools for adolescents used in Vietnam for research purposes have not been tested for validity and reliability. There are a number of possible reasons that may explain this finding.

First of all, the limitation in research funds might prevent researchers from conducting the cross-culture adaptation. Secondly, most lecturers, educators and researchers have not received any formal training about psychological assessment and cultural adaptation of assessment tools. Psychological assessment and evaluation have not been a compulsory subject in the Bachelor, Master and PhD training curriculum in Vietnam. There is also a lack of short courses related to this field across the country. If any courses are available, the cost of participation is too high to encourage participation. Thirdly, there are no reliable and easy to access guidelines how to conduct and evaluate the psychological assessment. Most guidelines are published in English and have not been translated into Vietnamese. Professionals and students who are competent in the English language, find it difficult to obtain professional resources as most universities in Vietnam are lack funding subscribe to the international online database. Fourthly, researchers who study overseas often remain in their host country and their obtained knowledge is not shared in their country of origin.

Inappropriate use of psychological assessment tools has been a serious problem in assessing mental health for adolescents in Vietnam. One of the initial steps to achieve the professional assessment is the adaptation of foreign tools. However, resources of guidelines and means to access them are limited. There should be a systematic approach that starts from a compulsory course for all psychology students at universities and funding to ensure better access to international databases.

## References


Amstadter, A., Richardson, L., Meyer, A., Sawyer, G., Kilpatrick, D., Tran, T., Acierno, R. (2011). Prevalence and correlates of probable adolescent mental health problems reported by parents in Vietnam. Soc Psychiatry Psychiatr Epidemiol, 46, 95-100.

Cao, V., Demombynes, G., Kwakwa, V., Mahajan, S., Shetty, S., Vu, D., & Trotsenburg, V. (2016). Vietnam 2035: toward prosperity, creativity, equity, and democracy - overview. Retrieved from Washington, D.C.: http://documents.worldbank.org/curated/en/2016/02/25967214/vietnam-2035-toward-prosperity-creativity-equity-democracy-overview

Critical Appraisal Skills Programme. (2017). CASP Checklist. Retrieved from http://www.casp-uk.net/referencing

General Statistics Office of Vietnam. (2005). SAVY I report. Retrieved from Vietnam:

General Statistics Office of Vietnam. (2008). SAVY II report. Retrieved from Vietnam:

General Statistics Office of Vietnam. (2016). Statistical Handbook of Vietnam. Retrieved from Vietnam:

Giang, K. B., Allebeck, P., Kullgren, G., & Nguyen, V. T. (2006). The Vietnamese version of the self reporting questionnaire 20 (SRQ-20) in detecting mental disorders in rural Vietnam: A validation study. International Journal of Social Psychiatry, 52(2), 175-184. doi:10.1177/0020764006061251

Haynes, S. N. (1998). Principles and practices of behavioural assessment with adults. Comprehensive Clinical Psychology, 4, 157-186.

Kaljee, L. M., Green, M. S., Zhan, M., Riel, R., Lerdboon, P., Lostutter, T. W., & Minh, T. T. (2011). Gender, alcohol consumption patterns, and engagement in sexually intimate behaviors among adolescents and young adults in Nha Trang, VietNam. Youth Soc, 43. doi:10.1177/0044118x09351285

Le, L. C., & Blum, R. W. (2015). Changes in and Challenges for Intentional Injury in Vietnam. Asia-Pacific Journal of Public Health, 27(2), NP1537-NP1548. doi:10.1177/1010539512448525

Le, M. T., Nguyen, H. T., Tran, T. D., & Fisher, J. R. (2012). Experience of low mood and suicidal behaviors among adolescents in Vietnam: findings from two national population-based surveys. J Adolesc Health, 51. doi:10.1016/j.jadohealth.2011.12.027

Le, M. T. H., Holton, S., Nguyen, H. T., Wolfe, R., & Fisher, J. (2016). Poly-victimisation and health risk behaviours, symptoms of mental health problems and suicidal thoughts and plans among adolescents in Vietnam. International Journal of Mental Health Systems, 10(1), 66. doi:10.1186/s13033-016-0099-x

Le, T. H. M., Holton, S., Nguyen, T. H., Wolfe, R., & Fisher, J. R. W. (2016). Victimisation, poly-victimisation and health-related quality of life among high school students in Vietnam: a cross sectional survey. Health and Quality of Life Outcomes, 14(155). doi:10.1186/s12955-016-0558-8

McKelvey, R., Davies, L., Sang, D., Pickering, K., & Hoang, C. (1999). Problems and competencies reported by parents of Vietnamese children in Hanoi. Child and Adolescents Psychiatry, 38(6), 731-737.

Nguyen, B. D. (2014). Post-trauma stress disorder in adolescents experiencing domestic violence. Vietnam Psychology Journal, 1(178), 51-58.

Nguyen, D., Dedding, C., Pham, T., Wright, P., & Bunders, J. (2013). Depression, anxiety, and suicidal ideation among Vietnamese secondary school students and proposed solutions: a cross-sectional study. BMC Public Health, 13, 1195-1195. doi:10.1186/1471-2458-13-1195

Nguyen, D. T., Dedding, C., Pham, T. T., & Bunders, J. (2013). Perspectives of pupils, parents, and teachers on mental health problems among Vietnamese secondary school pupils. BMC Public Health, 13(1046).

Nguyen, H. T. (2006a). Child maltreatment in Vietnam: prevalence and associated mental and physical health problems. (Doctor of Philosophy), Queensland University of Technology.

Nguyen, H. T. (2006b). Child maltreatment in Vietnam: prevalence and associated mental and physical health problems. Thesis for the degree of doctor of philosophy. Queensland University of Technology, Faculty of Health; 2006. (Doctor of Philosophy), Queensland University of Technology.

Nguyen, H. T., Dunne, M. P., & Le, A. V. (2009). Multiple types of child maltreatment and adolescent mental health in Viet Nam. Bulletin of the World Health Organization, 88(1), 22-30 29p. doi:10.2471/BLT.08.060061

Nguyen, H. T., & Nguyen, H. T. (2012). The effect of temperaments to anxiety among high school students. Vietnam Psychology Journal, 3(156), 24-33.

Pham, M. H., & Do, L. (2004). Psychology in Vietnam. The Psychologist, 17(2), 70-71.

Pham, T. T. B. (2015). Study burden, academic stress and mental health among high school students in Vietnam: Queensland University of Technology, 2015.

Pham, T. T. P. (2014). Behavioural disorder among students at Nguyen Binh Khiem private secondary school in Hanoi. Vietnam Educational Journal, 1(339), 20-21 and 29.

Pham, V. T. (2016). Aggressive symptoms in secondary school students. Vietnamese Journal of Psychology, 7(208), 50-59.

Stratton, K. J., Edwards, A. C., Overstreet, C., Richardson, L., Trinh, L. T., Lam, T. T., . . . Amstadter, A. B. (2014). Caretaker mental health and family environment factors are associated with adolescent psychiatric problems in a Vietnamese sample. Psychiatry Research. doi:10.1016/j.psychres.2014.08.033

Thai, T. T., Kim, X. L., Nguyen, D. N., Dixon, J., Sun, J., & Dunne, M. (2015). Validation of the Educational Stress Scale for Adolescents (ESSA) in Vietnam. Asia-Pacific Journal of Public Health, 27(2). doi:10.1177/1010539512440818

The World Bank. (2011). World development report 2011: Conflict, Security, and Development. Retrieved from Washington DC:

Tran, B. P., Nguyen, T. H., Truong, Q. T., Hoang, K. C., & Dunne, M. P. (2013). Factors associated with health risk behavior among school children in urban Vietnam. Global Health Action, 6, 1-9. doi:10.3402/gha.v6i0.1887

Tran, D. T., Tuan, T., & Fisher, J. (2013). Validation of the depression anxiety stress scales (DASS) 21 as a screening instrument for depression and anxiety in a rural community-based cohort of northern Vietnamese women. Bmc Psychiatry, 13(24).

Tran, T., Harpham, T., & Nguyen, T. H. (2004). Validity and reliability of the self-reporting questionnaire 20 items in Vietnam. Hong Kong Journal of Psychiatry, 14(3), 15-18.

Tran, T. N. (2015a). Anxiety in high school students and its association with self-esteem, study motivation and achievement. Vietnamese Journal of Psychology, 7(196), 45-55.

Tran, T. N. (2015b). The association between parental styles and behavioural-emotional disorders' symptoms in youth. Vietnamese Journal of Psychology, 4(193), 47-60.

Vu, K. (1997). Confucianism and its development in Vietnam. Vietnam: Social Science Publishing House.

## Testing and Psychodiagnostics in Ukraine

**Nariman Darvishov**
Department of Psychodiagnostics and Clinical Psychology of Taras Shevchenko National University of Kiev
Kiev, Ukraine

Psychological testing in Ukraine has long and nonlinear history of development.

At the beginning of the 20th century tests of the Russian scientific school were implemented in the territory of Ukraine. For example, the Psychological profiles of G. Rossolimo were used as a scale to measure general abilities and the Rybakov figures is an example of the test which were used to determine the spatial cognition. These and other scientific methods provided valid psychometric results. Sometimes some of them were competing with the leading ones in the world of psychometric practices at that time. Also some of the works on applied statistics in psychological testing (Mandryka, 1931) and on the link of the results of psychological testing with social factors (Syrkin, 1929) were published in Ukraine.

The first decade of the Soviet Union existence was marked by the development of pedology science. Ukraine as a part of the Soviet Union was also included in this process. The purpose of pedology was to measure the psychological characteristics of children in order to optimize the educational process. Pedological tests were useful in assessment of the common educational success and the mental age of students. However, pedology was banned at the government level due to the political reasons in 1936. Further practice of pedology was fraught with serious sanctions. In fact, that incident meant the termination of the quantitative measurement in psychology. Qualitative approach was an analogue for the tests applicable to a narrow range of tasks. It is a study of mental development through the performance of activities and the ability to perform which was a studied subject.

In the 1960s, psychological testing had a new lease of life on the territory of the Soviet Union, thanks to the new name - psychodiagnostics. Tests were gradually used in clinical psychology studies, age development studies, vocational guidance and vocational selection and in forensic psychology.

In 1978 the first monograph on psychodiagnostics “Psychodiagnostics of intellect and personality” was published by the authorship of Vadim Bleicher and Leonid Burlachuk.

In 1987 “General Psychodiagnostics” as the first in the USSR textbook on psychodiagnostics was published for students of psychology by the authorship of A. Bodalev and V. Stolin.

In 1989, L. Burlachuk and S. Morozov published the first in the USSR dictionary handbook on psychodiagnostics.

In 1992, the Department of Psychodiagnostics and Clinical Psychology at the Faculty of

Psychology of the Taras Shevchenko National University of Kyiv was formed on the initiative of L. Burlachuk. The department provides students with courses of psychodiagnostics, projective methods in psychodiagnostics, clinical psychodiagnostics, methods of forensic expertise. Students receive a practical task to develop their own testing methods as a part of the psychodiagnostics course.
Since 2007, Ukraine has a department of Giunti psychometrics – "OS Ukraine", as an associate member of the European group of tests publishers [1]. Giunti Psychometrics Ukraine is the official provider and partner of AUPA (All-Ukrainian Psychodiagnostic association) and maintains regular contacts with International Test Commission (ITC). The organization regularly conducts research to improve the quality of testing instruments.

One of the most significant projects in this area was the adaptation of MMPI-2 in Ukrainian language by a team of scientists: Leonid Burlachuk, Dmitry Korolev, Olga Morozova-Larina, Karine Malysheva, Natalya Zavyazkina, Alexander Vinogradov, Oleksandra Kryshovska, Olha Orel, Oleg Burlachuk [3]. In adaptation procedure there was a sample of 1178 people, including 178 patients with diagnosed psychopathology. Adaptation lasted for two years and was completed in 2014. The quality of Ukrainian adaptation was highly appreciated by the scientific collective of University of Minnesota and recognized as an official adaptation [4].

Large-scale studies with the use of test equipment were conducted in the field of judiciary and justice. For example, new adapted methods, including MMPI-2questionnaire, are used in the selection of the new national police service of Ukraine since 2015.

For the first time in Ukraine the Giunti psychometrics company has begun active adaptation of the WISC-IV – a most common technique and standardized method of children's intelligence measuring in the world. Also Giunti psychometrics involved more than 20 leading experts in four regions of the country to adapt the Wexler technique in Ukraine. As of today, the project is at the stage of collecting normative sampling - more than 1000 children from different cities and villages of Ukraine. For the April 2018, the adaptation of the test is on its final stage. The planned release date for WISC-IV is autumn 2018 [2].

In 2013 All-Ukrainian Psychodiagnostic Association (AUPA) was established at the initiative of professional psychologists involved in the development and adaptation of psychological tests. The structure of association also includes professors, graduate students and students of psychological faculties from all over Ukraine, professional psychologists with higher psychological education, which have consultative and psychotherapeutic practice. Common goal of association – the development of Ukrainian psychodiagnosis as one of important areas of applied psychology in Ukraine. Today AUPA presented in Kyiv and has offices in 16 regions [5].

Since 2018 AUPA has been developing a new advanced psychodiagnostic course for specialists.

We can observe a growing demand for the test tools and testing in Ukraine, both from the government institutions and from the private customers. Available tools require active modernization work. The number of testing research projects is increasing due to the AUPA association specialists' actions. It allows to establish professional communication

between specialists from all over the country and to integrate into the practice world of psychodiagnostics.

**References:**


1. URL: http://etpg.org/members_associate_oz.html
2. URL:https://giuntipsy.com.ua/about/
3. URL:https://giuntipsy.com.ua/clinical/mmpi-2/
4. URL:https://www.upress.umn.edu/test-division/translations-permissions/available-translations/ukrainian
5. URL:http://vpa.org.ua/pro-nas/


## Enhancing Career Guidance and Counseling Practice in Malaysia with Psychometrically-Sound Assessment Tools


**Sabrena Arosh**
University of Malaya
Kuala Lumpur, Malaysia


Malaysia is a relatively young South-East Asian nation north of Singapore and south of Thailand that succeeded in achieving independence a scant 61 years ago after many years of British rule. As such the structure of the educational system was largely seen as a replication of the British system although it has undergone much change since the nation's Independence. Aside from the general use of achievement tests to measure student progress, assessment in the field of career guidance and counseling has come to the forefront of our nation's educational planning since the early 1970s.

However, the use of West-oriented tests in this area has been noted needing further research due to lack of local norms while only a select few being translated and re-normed (Pope, Musa, Singaravelu, Bringaze, & Russell, 2002). Most studies in Malaysia focused on interest testing as seen in the work of Amla Mohd. Salleh (2010). Drawing from the pioneering work of Awang (1976, as cited in Amla H. M. Salleh, 1984) that translated Holland's Vocational Preference Inventory to Malay for administration among Malay secondary school students, Amla H. M. Salleh (1984) conducted a similar translation on Holland's Self-Directed Search and expanded the validation of its psychometric properties across different groups in Malaysia (Amla Mohd. Salleh, 2010).

In recent years, aptitude testing has slowly become a point of focus with the implementation of school-based assessment (Pentaksiran Berasaskan Sekolah, PBS). Originally proposed in 2007, PBS included additional forms of assessment aside from the typical central examinations, which included psychometric tests to assess students' abilities and interests as well as their readiness for learning (Ong, 2010). According to the proposed system, these assessments would be put in place up to the lower secondary level, but the upper secondary level would solely focus on centralized examinations (Ong, 2010).

The PBS plans finally came to fruition this year as Primary Year 6 students who sat for the 2017 Primary School Assessment Examination (Ujian Penilaian Sekolah Rendah, UPSR) received the Primary School Assessment Report which included assessment results for their classroom, psychometric, sports, physical activity, and curriculum components of PBS (Azura Abas & Hashini Kavishtri Kannan, 2017). However, such psychometric assessment merely encompassed aptitudes in

areas such as music, language, mathematical logic, kinaesthetic intelligence, and naturalistic intelligence (Azura Abas & Hashini Kavishtri Kannan, 2017). It was also unclear how these assessment results would be used in career exploration and guidance counseling for students.

The inclusion of these components under PBS was continued under the Malaysia Education Blueprint for the period of 2013-2025 (Ministry of Education, 2013). While the blueprint stated that aptitude testing would focus on measuring thinking and problem-solving skills (Ministry of Education, 2013), it is unclear if this will be used to inform the education or career counseling processes. In addition, Ong (2010) has outlined the challenge faced by the Ministry in moving towards holistic assessment including the development of psychometrically-sound tools after a longstanding historical focus on achievement testing, as well as proper communication and delivery of results. Moreover, other initiatives have also been undertaken by the country's private sector. For instance, HELP University, a private tertiary institution, developed a test battery meant for career exploration for students transitioning from secondary to college level (Chong et al., 2016; Darwishah et al., 2016; Mamauag, Tai, Arosh, Teo, & Ooi, 2016). The test battery, which consists of measures of career readiness, personality, employability skills, interests, and aptitude, is the first of its kind comprehensively developed within the Malaysian cultural context. This initiative holds great promise such that with the development of such localized assessment tools, better delivery of high quality, evidence-based services to students can happen, thus enhancing the career guidance and counseling practice in Malaysia.

## References


Amla H. M. Salleh. (1984). An investigation of the reliability, validity, and translation of Holland's Self Directed Search for utilization by a Malaysian population (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (ProQuest document ID: 303311846).

Amla Mohd. Salleh. (2010). Pendidikan kerjaya dan pembangunan modal insan. Selangor: Penerbit UKM.

Azura Abas & Hashini Kavishtri Kannan. (2017, November 21). Those picking up their UPSR results on Thursday, brace yourselves for something different. New Straits Times. Retrieved from https://www.nst.com.my/news/nation/2017/11/305854/those-picking-their-upsr-results-thursday-brace-yourselves-something.

Chong, N., Arosh, S. G., Mamauag, M. F. M., Teo, S. W., Tai, Y. S., & Ooi, H. J. (2016, July). Defining constructs of the Five Factor model using a Malaysian context: Towards developing an assessment tool for personality. Paper presented at the 31st International Congress of Psychology, Yokohama, Japan. Abstract retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ijop.12345/epdf.

Darwishah, N., Chong, N., Mamauag, M. F. M., Tai, Y. S., Arosh, S. G., Teo, S. W., Ooi, H. J., Wong, I., & Pang, W. T. (2016, July). Defining 21st Century Skills in the Malaysian context: Towards developing an assessment tool for employability and career readiness. Paper presented at the 31st International Congress of Psychology, Yokohama, Japan. Abstract retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ijop.12345/epdf.

Mamauag, M. F. M., Tai, Y. S., Arosh, S. G., Teo, S. W., & Ooi, H. J. (2016, July). Establishing the psychometric qualities of the Multiple Aptitude Test using the Rasch model. Paper presented at the 31st International Congress of Psychology, Yokohama, Japan. Abstract retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ijop.12345/epdf.

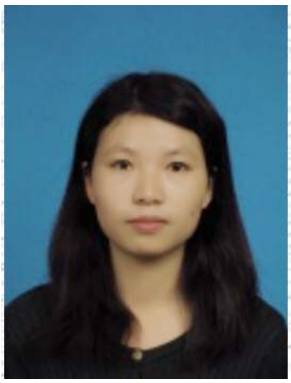
Ministry of Education. (2013). Malaysia Education Blueprint 2013-2025 (Preschool to Post-Secondary Education). Putrajaya: Ministry of Education Malaysia.

Ong, S. L. (2010). Assessment profile of Malaysia: High-stakes external examinations dominate. Assessment in Education: Principles, Policy, and Practice, 17(1), 91-103. doi: 10.1080/09695940903319752.

Pope, M., Musa, M., Singaravelu, H., Bringaze, T., & Russell, M. (2002). From colonialism to ultranationalism: History and development of career counseling in Malaysia. The Career Development Quarterly, 50(3), 264-276. doi: 10.1002/j.2161-0045.2002.tb00902.x.

## Item Response Theory in Mainland China

**Fen Ren**
School of Education and Psychology
University of Jinan, Shandong, China



As one key part of the modern testing theories, item response theory (IRT) has been developing extremely fast to achieve remarkable attention in mainland China during past decade (Lord, 1980).

In between 2000 and 2018, there were a total of 1,528 papers published in the field of psychology containing the keyword of ”item response theory”, which could be found in the Baidu scholar searching engine. Among these papers, 233 of them were from the Chinese social science Citation Index (CSSCI) database, 54 from the Chinese Science Citation Database (CSCD), and 1066 related to clinical medicine. Meanwhile, there are several top Chinese journals publishing papers on issues concerning IRT methodology and application, such as ‘Acta Psychologica Sinic’, ‘Chinese exams’, and ‘Chinese Journal of Clinical Psychology’. On the other hand, several popularly used text books on IRT have also been published since 1990s (Lin,1990; Xu,1992; Yu,1992).

The most salient change could be observed on the application of IRT in empirical studies. For instance, an increasing number of researchers have been applying IRT techniques to address diverse topics in various academic fields. Moreover, great contributions have been made by scholars to methodological development and promotion of IRT methods, which in turn popularize this theory in many areas. Two universities, the Beijing Normal University and the Jiangxi Normal University, are leading the fast spread of this modern theory.

For the most widely used areas, education area, the topic mainly focus on cognitive diagnosis model, such as unified model, fusion model, deterministic in put, noisy and gate model as well as noisy inputs, deterministic, and gate model. But for psychological area, multidimensional item response theory are used to estimate the relationship between dimensions as well as parameters.

In July 2015, the 80th International Meeting of the Psychometric Society (IMPS), one of the most influenced events in psychological measurement, was held in Beijing, hosting by the Beijing Normal University. This was the first time that IMPS had ever been held in mainland China, suggesting the important role of Chinese scholars in this filed. More than 500 participants from all over the world joined the conference to share their most updated research on psychological and educational measurement, including model development and real data application (Mu Honghua, Ting Wang, & Jianlin Wang, 2015 ). It is not surprising that the meeting placed a significant impact on researchers, students, and educators in psychology, especially those majored in quantitative methods.
Besides, using IRT to analyze test or exam data could also bring financial benefits. For example, the first company whose main

business was providing professional consultation to and technical supports on tests in China was established in 1997.

In sum, IRT has been developing dramatically in China, and achieved certain success and acknowledgment, particularly in the areas of psychology and education. With joint efforts of scholars and practitioners, a bright future would definitely be guaranteed.

### References:


Lin He. (1990). Item Response Theory. Hubei Education Press.

Lord, F. M. (1980). Applications of item response theory to practical testing problems , 1, 274.

Mu Honghua, Ting Wang, & Jianlin Wang. (2015). Review of The 80th International Meeting of the Psychometric Society(IMPS). Educational Measurement and Evaluation (12),11-15.

Jiayuan Yu. (1992). Item Response Theory and its Applicaiton. Jiangsu Education Press.

Zuweu Xu.(1992). Item Response Theory and its Application. East China Normal University Press.


## Psychometric testing within occupational settings in a multicultural, developing country


**Angela (Lee) Marsburg**

Stellenbosch University, South Africa


Testing and assessment has been a source of recurring social controversy within South Africa's history (Laher & Cockcroft, 2014; Theron, 2007). There has subsequently been much debate pertaining to how best to protect an already skeptical, untrusting and suspicious South African public (Pitoniak & Yeld, 2013), while additionally improving the credibility and integrity of psychological testing, which was previously used as a means to legitimize segregation and division during the apartheid era (Donald, Thatcher, & Milner, 2014). This article aims to provide a brief overview of the current challenges facing test use, specifically within organisational contexts in South Africa.

### The current South African context

The heterogeneity of the population, currently standing at over 55 million people (United Nations, 2017), with 11 official languages, from various cultural and religious backgrounds, makes cross-cultural aspects, such as fairness and bias within testing and assessment, very salient. Reviewing standards of testing practices globally, there are typically two approaches (that are not mutually exclusive) to mitigate the risk of testing causing harm, while advancing the benefits of test use, that according to Bartram (2012) are anchored in education (focusing on the competence of test users and the quality of the instrument) and restriction (controlling access to test use). To this end, the Employment Equity Act (EEA) 55 of 1998 (Department of Labour, 1998) served the function of setting standards for the use of psychometric testing within South African occupational settings, for over a decade. The EEA (Department of Labour, 1998, p. 16) stipulates that "psychological testing and other similar assessments of an employee are prohibited unless the test or assessment being used a) has been scientifically shown to be valid and reliable, b) can be applied fairly to all employees and c) is not biased against any employee or group." Meeting these requirements alone does not however ensure the complete negation of potential harm to test takers.

The Health Professions Council of South Africa (HPCSA), which serves an oversight function for all health professions in the country, therefore additionally oversees psychological

test use via the Psychometrics Committee of the Professional Board of Psychology. Tests are reviewed by the Psychometrics Committee and classified as either psychological tests or tests that can be used by other professionals. Following a restriction approach (Bartram, 2012), testing that measures a psychological construct, as defined in the Health Professions Act, 56/1974 (Republic of South Africa, 2008), is therefore limited to registered psychological practitioners with the relevant empirical and theoretical knowledge, and professional training (Department of Health, 2009; Republic of South Africa, 2008; Schmidt, 2006).

### Recent legislative developments

The diverse nature of our population, and the acknowledgement of balancing the value of assessments against potential misuse, has led to a highly regulated assessment environment, when compared internationally. Furthermore, 2016 saw the promulgation of a fourth clause in the amended EEA (Republic of South Africa, 2014). This clause stated that psychological testing and other similar assessments must be *certified* by the HPCSA, or a third party that they nominate. This amendment, and the subsequent publishing of a gazetted list of classified tests by the HPCSA, meant many stakeholders found themselves non-compliant overnight (Association of Test Publishers, 2015). The amendment additionally failed to offer protection to test takers from unjustified discrimination (the purpose of the Act), in that it did not address issues around criterion inferences derived from sound valid, reliable, predictor measures that can still contain systematic group-related error and are subsequently predictively biased.

This amendment was strongly contested, by key stakeholders within the testing and assessment industry, and subsequently deemed null in void on procedural grounds (High Court of South Africa, ATP v President, 89564/14, 2017), maintaining the status quo. Practitioners must therefore still provide evidence of the quality of the instruments in use, the cross-cultural applicability as well as the appropriateness of the norms used, and the Psychometrics Committee will proceed with the current classification system. The amendment however brought about an urgency to discussions over the last 10 years, regarding the alignment of the current classification process with international standards on test review. In this regard, the proposed re-imagining of the current classifications system is based on the European Federation of Psychologists Associations (EFPA) standards for psychological assessment (HPCSA, 2015).

### Testing practices in South Africa

Within the occupational context, testing predominantly occurs in English (accepted as a business language), even though most South African's are bi-or-multilingual and do not speak English as a first language. Ethnicity and language variables are therefore of particular concern when evaluating an instrument for the presence of bias. Previously, there was a strong reliance on imported tests. Recently, there has been a call for measures that are not only translated and adapted, but rather developed using a combined emic-etic approach, that integrates indigenous and universal dimensions of culture (Odendaal, 2016). This is in a bid to capture the nuances of South Africa's diverse culture and languages (multilingualism) in personality assessment, examples include the South African Personality Inventory and the Basic Traits Inventory.

A search for an intellectually honest response to the legacy of apartheid, and its subsequent negative influence on the crystallized intelligence of certain population groups, within South Africa, also lead to an emphasis on affirmative development, and the subsequent use of psychological assessment to ascertain

levels of learning potential among these populations, to identify suitable developmental candidates. A need still exists however for an intellectually honest solution to the challenge of adverse impact, via predictive bias in selection. It can be argued that changing the measure, used across groups, does not provide this solution, due to the negative influence of apartheid, and subsequent lack of development, in terms of core workplace competencies, within specific populations. The solution lies in the realisation that fundamentally talent is not related to gender, race or creed. Fundamentally, the solution therefore is to be found in identifying high potential individuals in disadvantaged populations and affording them the development opportunities that they were denied, and thereby attempt to close the gap in the criterion distributions across groups.

## Looking to the future: challenges and opportunities

Unfortunately, legislation and classification does not, in-and-of-itself, ensure that testing, and subsequent decisions made using tests, are fair. Currently, the scientific merit of psychological tests, and other similar assessments (e.g. Interviews, climate surveys), provides merely one conceptualisation of validity (Schmidt, 2006). The current technical rationality that underpins the prevailing concept of validity captured by the EEA, according to Schmidt, "does not account for the normative dimensions inherent to the psychological act of assessment" (Schmidt, 2006, p.60). It can be argued that the inferences drawn from valid, reliable and unbiased measures, especially in high stakes decision making situations in occupational settings, are equally as important in determining whether the decision discriminates fairly or unfairly between employees.

There are still numerous pertinent unanswered questions, and uncertainty is a challenge that we as practitioner's face. These pertain to, not only the question of classification versus certification, but also to what constitutes the defining characteristics of "other similar assessments" (Department of Labour, 1998, p. 16), the types of tests that will not require classification and why, appropriate punitive action if non-classified tests are used, and how the regulatory bodies plan to aid the plight of local test developers, by providing a transparent procedure and criteria for classification. Furthermore, as a developing country, a related and serious question is whether we have the required resources and capacity to facilitate the complete overhaul of the procedures and processes that are necessary. It is hoped that through the continued collaborative efforts, and inputs from both local and global stakeholders, these questions will be answered, and a solution that is tailor made to protect the South African public and maintain the integrity of testing and assessment will be sought.

## References


Association of Test Publishers. (2015). Updated position statement on the Amendment Section 8 of the Employment Equity Act No 55 of 1998.

Bartram, D. (2012). Concluding Thoughts on the Internationalization of Test Reviews. *International Journal of Testing*, *12*(2), 195–201.

Department of Health. Health Professions Act, 1974: Regulations defining the scope of the profession of psychology, No. 756, J Government Gazette (2009).

Department of Labour. (1998). Employment Equity Act No 55, 1998. *Government Gazette*.

Donald, F., Thatcher, A., & Milner, K. (2014). Psychological assessment for redress in South African organisations: Is it just? *South African Journal of Psychology*, *44*(3), 333–349.

Helath Professions Council of South Africa. (2015). *Report on the Stakeholder engagement of the Psychometrics Committee of the Professional Board of Psychology.* Emperors Palace, Kempton Part, 5 March.

High Court of South Africa. (2017). *Judgement in matter between ATP and president of RSA, Minister of labour and HPCSA*.
Laher, S., & Cockcroft, K. (2014). Psychological assessment in post-apartheid South Africa: The way forward. *South African Journal of Psychology*, *44*(3), 303–314.
Odendaal, A. (2016). Emerging trends in the development, control and use of psychological tests in South Africa. In *10th ITC Conference*. Vancouver, Canada.
Pitoniak, M. J., & Yeld, N. (2013). Standard Setting Lessons Learned in the South African Context: Implications for International Implementation. *International Journal of Testing*, *13*(1), 19–31.
Republic of South Africa. Health Professions Act 56 of 1974, Goverment Gazette § (2008).
Republic of South Africa. (2014). Employment Equity Amendment Act No. 47 of 2013. *Government Gazette*, *583*(37238), 2–22.
Schmidt, C. (2006). Validity as an action concept in IO psychology. *SA Journal of Industrial Psychology*, *32*(4), 59–67.
Theron, C. (2007). Confessions, Scapegoats and Flying Pigs: Psychometric Testing and the Law. *SA Journal of Industrial Psychology*, *33*(1), 102–117.
United Nations. (2017). World population prospects: The 2017 revision key findings and advanced tables. *World Population Prospects*, 1–46.

## Don't Forget: ITC Guidelines on the ITC Website

The ITC guidelines bear directly on furthering the goals of the ITC. Six projects have produced guidelines that have gained wide international acceptance. These are:

1. **The ITC Guidelines on Adapting Tests**
2. **The ITC Guidelines on Test Use**
3. **The ITC Guidelines on Computer-Based and Internet-delivered Testing**
4. **The ITC Guidelines on Quality Control in Scoring, Test Analysis and Reporting of Test Scores**
5. **The ITC Guidelines on the Security of Tests, Examinations, and Other Assessments**
6. **The ITC Guidelines on Practitioner Use of Test Revisions, Obsolete Tests, and Test Disposal**

You can download the Guidelines by accessing their pages on the ITC website.

## Festival international de jazz de Montréal

http://www.montrealjazzfest.com/

The Festival International de Jazz de Montréal has been synonymous with a passion for music for nearly 40 years. Every year for 10 days, the French-speaking metropolis of North America becomes the venue where fans of all types of jazz-related music rub shoulders with aficionados of jazz in its purest form. All on a unique site designed to meet festival goers' every need, right downtown in an area off-limits to car traffic! Montreal is without a doubt the true heartbeat of Planet Jazz!

## A Brief Update on the ITC Book Series

Neal Schmitt, Chair

The first two volumes in the ITC book series published by Cambridge University are now available. The first book authored by Dragos Iliescu is titled “Adapting tests in linguistic and cultural contexts. John Scott, Dave Bartram and Douglas Reynold’s book titled “International applications of web-based testing: Challenges and Opportunities” has just become available. The volume by William Schmidt and his colleagues titled “Measuring opportunity: Insights from international large scale assessment” should be available this summer. Look for discounts on these volumes at the ITC conference in Montreal.

Craig Wells continues working on “Assessing measurement invariance for applied researchers” and Maria Elena Oliveri and Cathy Wendler are editing a volume titled “Higher education admission practices: An international perspective.”

We are still hoping to enlist someone interested in doing books on the history of testing internationally, personality testing, and test security challenges when working internationally. Anyone interested in these topics or others is invited to email us and explore possibilities. I will describe the process of book proposal and writing in an effort to enlist new authors interested in addressing topics of interest to an international audience.

## Book Review: The Wiley Blackwell Handbook of The Psychology of Recruitment, Selection and Employee Retention

Edited by
Harold W. Goldstain, Elaine D. Pulakos, Jonathan Passmore and Carla Semendo.
Series Editor Jonathan Passmore
ISBN 978 1 118 97269 4 Wiley Blackwell: NJ

This 568 page beautifully produced handbook is likely to become an essential addition to any organisational library. As a resource for anyone looking for a summary of latest trends, research and issues, it would be invaluable. The Editors have pulled together 24 chapters, covering recruitment, selection and retention in details, with contributions from a total of 57 respected professionals in the field. As a result, there are few areas left uncovered, although several have been flagged as needing more research.

The book is organised into three sections: the first six chapters are on Recruitment, the following 14 chapters deal with various aspects of Selection, and the final 4 chapters explore the question of Retention. Section 1 begins with an outline of the overall content, and then goes on to explore the specifics of job analysis, global recruiting, aspects of the selection process from the point of view of the candidate, issues to do with attracting talent, and the ethics of recruitment and selection. The material covered ranges from a useful, evidence-based "how-to" guide to job analysis, to theoretical models of applicants' reactions to the selection process, including the moral and ethical aspects of looking at candidates' social networking to influence selection.

Section 2 is by far the largest section of the book, and covers the selection process in detail. It looks at the use of ability tests and personality questionnaires, interviewing, SJTs, biodata and simulation exercises, in each case reviewing the literature and highlighting issues and recommendations arising from it. Chapters then go on to examine the research evidence surrounding online selection, gamification and virtual team selection. Leader development, team assessment and talent management are also covered, and the section ends with a chapter on cultural differences followed by one on legislative pitfalls and fairness consideration. Each chapter covers both established research and more recent developments, and highlights areas which need more research as well as those where research has led to clear recommendations for good practice. While individual chapters inevitably reflect the individual interests and concerns of their authors, the extensive range of this section results in a comprehensive coverage of the selection process, useful both for practitioners and researchers.

The third section deals with the question of retention, and contains a useful review of employee turnover and retention strategies, including internet and international contexts. The discussion covers dynamic models of turnover and explores what is needed to generate positive work climates. Maximising human capital means managing and retaining talented individuals, which can make all the difference to organisational success, and chapters in this section explore issues of leadership style, organisational climate and culture, reward systems, and identification. They also highlight a number of key gaps in research - I would recommend any chapter in this section to a graduate student looking for a relevant research project. The final chapter in this section discusses issues of work-life balance, looking at conflicts and moderators of potential conflict, support policies and how organisations can assist employees in maintaining a positive balance.

What this book provides the student looking for a research area, then, is an overview of relevant research and the identification of gaps in our knowledge, which could usefully be studied. What it offers the professional in the field is much more than that. I found it an invaluable guide to evidence-based practice. It outlines findings, discusses issues, and clarifies principles, and in my view, no organisational library or consultancy should be without it. It's not cheap (slightly less online) but clearly written and, given its breadth and depth of coverage, well worth the cost.

*Submitted by incoming TI Editor Nicky Hayes*

**Call for Papers and Announcements:**
***Testing International (TI)***

**Deadline for December 2018 issue: November 15, 2018**

*TI* is the newsletter of the International Test Commission, and disseminates information about national / international assessment projects and initiatives, test developments, recently published books / articles, upcoming conferences and workshops, and topical issues in the field of testing and assessment to the international community.

Please contact Dr. Nicky Hayes, with proposals, announcements, and brief papers!

***tieditor[at]intestcom.org***

# Answering open ended questions: Does it matter if test takers have a choice of which questions to answer?

**John J. Barnard**
EPEC Pty Ltd, Melbourne, Australia &
University of Sydney, Australia
**L. Davies**
University of Sydney, Australia
**N. G. Chiavaroli**
University of Melbourne, Australia
**M. Trigg**
EPEC Pty Ltd, Melbourne, Australia

## Abstract

Examinations remain one of the most popular methods for assessing knowledge. There are various questions formats used in exams and one of the most utilised formats includes short-answer questions (SAQs) (Mullen, K & Schultz, M 2012). SAQs require examinees to provide a written answer of varying length to a prompt and depending on the exam, candidates can either be required to respond to all questions or select some questions to answer. In the instance where all questions are compulsory, it may be fair to compare the overall scores of the candidates directly. If however, candidates are allowed to answer different combinations of questions, the common practice of comparing total scores may in fact disadvantage or advantage particular candidates due to the varying difficulty of the questions selected and the level of knowledge required to obtain a certain score within a question.

This study includes two types of analysis of the examination results of 118 candidates; classical analysis based on Classical Test Theory (CTT) and Rasch analysis. Candidates were required to answer any four of eight SAQs. When classical analysis was applied, some candidates were advantaged and some disadvantaged due to the varying difficulty of the questions they selected as well as the inconsistent levels of knowledge required to obtain different scores within questions themselves. It was concluded that applying Rasch analysis was the most appropriate methodology to obtain accurate results for candidates in order to compare performance across any combination of four questions from a total of eight SAQs.

## Introduction

Although multiple-choice questions (MCQs) have become an increasingly popular question type in examinations for assessing knowledge, short-answer questions (SAQs) are still being used widely to determine understanding (Hudson & Treagust, 2013; Ventouras, Triantis, Tsiakas & Stergiopoulos, 2010; Wainer & Thissen, 1993). When SAQs are included in exams, some institutions allow candidates to select which questions they would like to answer from a given set of questions and others require candidates to answer all questions.

When offered an opportunity to select which questions to answer, a candidate will likely select the questions they consider to be the 'easiest' in order to maximise their overall score in the exam (Barnett-Foster & Nagy, 1996). However, the perceived difficulty of a particular question may be subjective to each candidate. Through appropriate analysis, it is possible to determine the true difficulty of such questions, which allows for the comparison of scores across questions and as such, the real performance of candidates.

In this article, the overall (total) score calculation, using classical (traditional) analysis is compared to deriving scores based on

Rasch analysis, which takes question difficulties and scoring structures within each question into account to determine overall performance. Results and associated implications are discussed.

## Study

Data from 118 candidates from a medical college in Australia who sat a high stakes qualification exam was obtained for analysis. Candidates were required to answer any four of eight questions. Each question was scored independently by two subject matter experts on a scale from 0 to 10 and the consensus score of the experts was taken as the final score for each question.

The aim of the study was to investigate whether the choice of questions made any difference to the overall result obtained by candidates.

## Analysis and Results

The performance of the candidates was determined in two different ways. Firstly a total score was calculated for each candidate using classical analysis. Secondly, the relative difficulties of the questions were determined using Rasch analysis after which the performance of each candidate was calculated and expressed as scores (Trigg, Barnard, Pham & Devitt, 2016; Petrillo, Cano, McLeod & Coon, 2015; Magno, 2009).

Based on classical analysis, a mean total score of 20.44 (SD 5.62) was found for the 118 candidates with total scores ranging from 8 to 37 out of 40.

A total score of 21 was the most frequent score (mode) and combinations of the 11 candidates who scored 21 in their question choices are shown in Table 1 below. The mean for each question for these 11 candidates is shown below the individual question scores and the mean and standard deviation for each question for the 118 candidates is also shown at the bottom of the table.

Table 1: Scores by question for the 11 candidates who scored 21 overall

| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 |
|---|---|---|---|---|---|---|---|---|
| **Candidate One** | 5 | 4 | | | | 6 | | 6 |
| **Candidate Two** | | 5 | | | | 4 | 5 | 7 |
| **Candidate Three** | 5 | | | | | 3 | 5 | 8 |
| **Candidate Four** | 6 | | | | 4 | | 6 | 5 |
| **Candidate Five** | 4 | 7 | | 4 | | | | 6 |
| **Candidate Six** | 7 | 5 | | 4 | | | | 5 |
| **Candidate Seven** | | 3 | 7 | | | | 5 | 6 |
| **Candidate Eight** | 7 | | | 3 | | 5 | | 6 |
| **Candidate Nine** | 6 | 6 | | | 4 | | 5 | |
| **Candidate Ten** | | 5 | | 5 | | | 5 | 6 |
| **Candidate Eleven** | | 7 | 4 | 2 | | | | 8 |
| **Mean** | **5.7** | **5.3** | **5.5** | **3.6** | **4.0** | **4.5** | **5.2** | **6.3** |
| | | | | | | | | |
| **All 118 mean** | 5.13 | 5.52 | 3.53 | 3.69 | 5.12 | 4.98 | 4.60 | 5.96 |
| **All 118 SD** | 1.83 | 1.81 | 2.41 | 1.99 | 1.51 | 2.07 | 2.12 | 1.67 |

From Table 1, both sets of means indicate that Question 8 was the 'easiest' question (having a mean score of 6.3 out of 10 for the 11 candidates) and significantly easier than, for example, Question 4 (having a mean score of 3.6 out of 10 for the 11 candidates).

The summary statistics presented in Table 1 above are typically reported when CTT is used as the underlying measurement model (Barnard, 2012; Crocker & Algina, 1986). However, these statistics are calculated on the basis of several assumptions. For example, it is assumed that the difference between a score of 3 and a score of 4 is the same as the difference between a score of 8 and a score of 9 (i.e. that there is a constant difference of 1). Calculating a total score using CTT is furthermore based on the assumption that the questions are of equal difficulty. This assumption is not realistic as can be seen in the results shown in Table 1 where there are eight different means for eight different questions. Many other shortcomings including; item statistics being sample dependent, measurement precision of individuals not being available and question difficulties and candidates' performance not being located on a common scale are well documented in the application of CTT (Magno, 2009; Hambleton & Swaminathan, 1985).

Rasch modelling was used as the second method to determine the performance of the 11 candidates based on the questions they answered. Rasch measurement theory purports to overcome the CTT issues by estimating candidate abilities and question difficulties separately on an interval level scale through logarithmic transformations (Tor & Steketee, 2011; Wright, 1977). Instead of scores, these estimates are expressed in log odds (logits) units, which usually range from -3 to 3 although higher and lower values are possible (Bond & Fox, 2013; Barnard, 2012). In Rasch models, the probability of a candidate of a certain ability to answer a question of a certain difficulty correctly is estimated from the data (Bond & Fox, 2013).

The family of Rasch models are based on the idea that data must conform to some reasonable hierarchy of 'less than/more than' on a single continuum of interest. (Barnard, 2012; Andrich, 1988; Andrich, 1978). Rasch rating scale analysis is more complex than modelling dichotomously scored questions, but the methodology is similar (Petrillo et al, 2015). Like dichotomous questions, each question's relative difficulty is estimated, but, in addition, the pattern of the scale categories to yield a rating scale structure is also established. Thus, each polytomous question has an overall difficulty estimate, but also has a series of thresholds for each score or category on the scale. Graphically this can be represented as shown in Figure 1.

Figure 1: A Polytomous Question

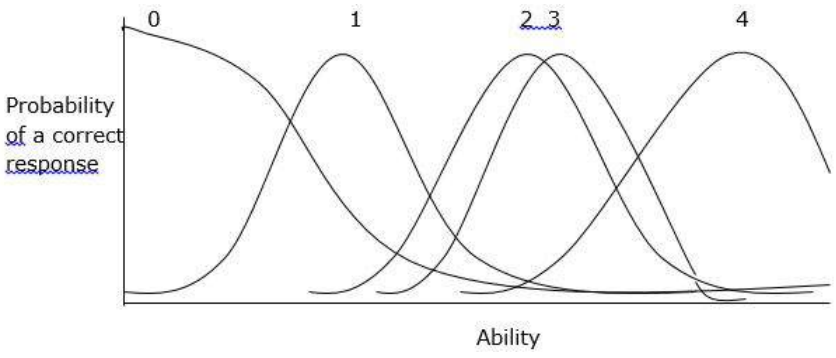


Figure 1 illustrates the ability (increasing from left to right) estimated to achieve certain scores. In this example significantly more ability is required to improve from a score of 1 to a score of 2 whereas the difference in ability necessary to score 2 is not much less than that required to score 3. In this case there is an argument for collapsing these two categories into a single scoring category.

Rasch analysis assumes that the data fit the model. This is verified by calculating fit statistics and exploring the targeting of the questions to the candidates, i.e. how well the difficulties of the questions match the abilities of the candidates. The latter can be done through plotting the distribution of the candidate abilities against the distribution of the question difficulties.

A Rasch calibration was done to estimate candidate abilities and question difficulties from the data. Figure 2 shows the distribution of the ability estimates (red bars) of all 118 candidates relative to the overall question difficulties (blue bars).
However, each question has 11 different score categories (0 to 10), which provide more detail about scoring in the question than the single overall question difficulty. Figure 3 below shows the scoring structure for all the question categories for each question.

Figure 2: A map of the distribution of candidate ability estimates against question difficulty estimates

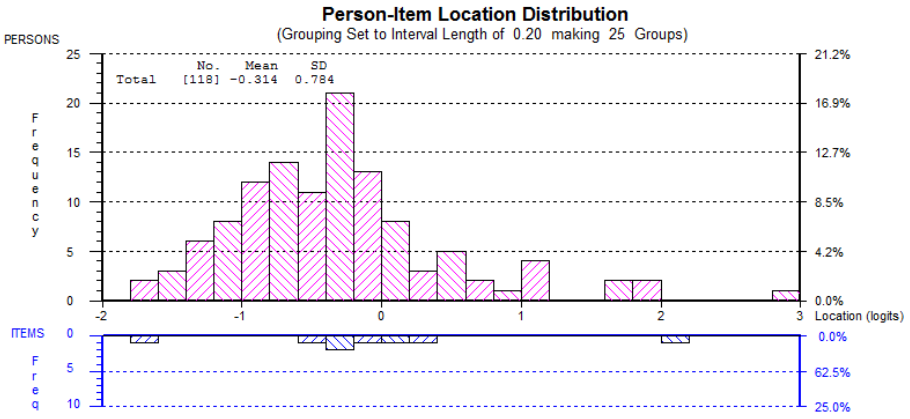



Figure 3: A map of the distribution of candidate ability estimates against question category difficulty estimates

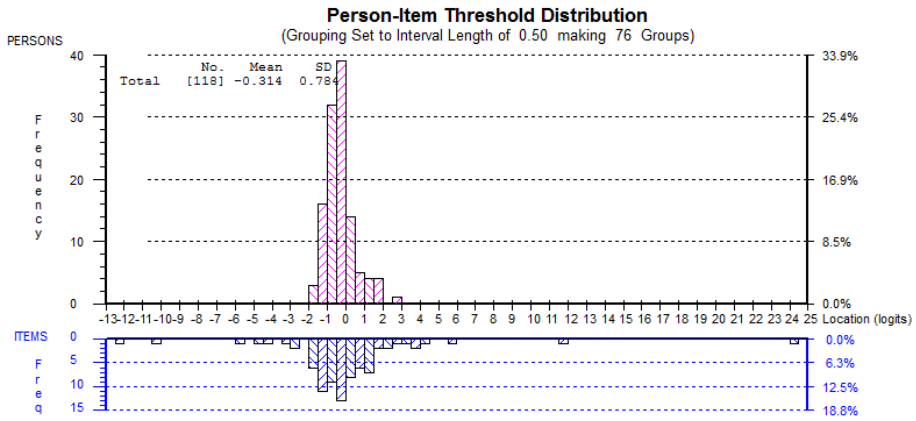


Figure 3 confirms a good match between the candidate abilities and the question category difficulties with some extremes. Note, for example, that it was very difficult to score in the highest category (10) in one of the questions, but also very easy to score in the lowest category (1) in one of the questions. If a certain question is the most difficult overall it doesn't necessarily mean that it is the most difficult to score in the highest category of that question. Likewise for the easiest question and the lowest score category. This is the essence of looking at the difficulties to score in the different score categories of the questions.

A typical Rasch calibration sets the mean question difficulty at zero logits and conclusions can be drawn from the ability estimates relative to the mean of the question difficulties. For this data a mean ability of -0.314 logits (SD 1.049) was found which suggests that the candidates found the questions 'slightly difficult' on average as the mean ability is less than the mean question difficulty of zero logits.
Table 2 summarises the resulting question difficulties and their measurement precision expressed in terms of standard errors.

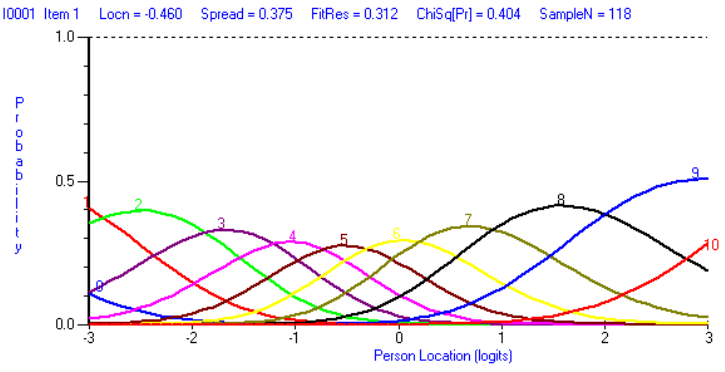
This analysis shows that Question 2 was the 'easiest' question overall and that Question 5 was the 'most difficult' question overall; in contrast to what was found in the classical analysis (where for example, question 8 was determined to be the 'easiest'). This difference can mainly be ascribed to the Rasch analysis accounting for the differences between scores within the questions whilst CTT assumes that all these differences are constant.

The different scores within a question can be depicted in category probability curves (CPC). The CPC of Question 1 in Figure 4 shows the desired pattern where there is a clear distinction between scoring in different categories.

Table 2: Question difficulties expressed in logits

| Question | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **Diff** | -0.460 | -1.655 | 0.130 | -0.026 | 2.132 | -0.200 | 0.305 | -0.226 |
| **SE** | 0.083 | 0.080 | 0.129 | 0.117 | 0.200 | 0.087 | 0.087 | 0.066 |

Figure 4: Category probability curves of Question 1



Unlike the steady progression from one scoring category to the next as shown in Figure 4 above, the CPC of Question 5 in Figure 5 suggest some clustering around the middle and also that 'very high' abilities are required to score in the higher categories. The difficulty of scoring in the high categories in Question 5 can be seen by looking at the higher values on the horizontal axis. For example, at 3 logits the most likely score is 7 for question 5 whilst it is most likely a score of 9 in Question 1 (see Figure 4 above).

With Question 5 being the most difficult question overall, it is also evident from Figure 5 that questions 5 was also the most difficult question to obtain a score of 10 in, but it was also the easiest question in which to score above zero. The most likely progression in Figure 5 is from a score of 2 (green curve) to 3 (purple curve) to 6 (yellow curve) to 7. Scores of 4 and 5 were thus never the most likely scores for any ability in question 5.

The relative difficulty to score in different categories in the questions can also be shown in what is referred to as a threshold map. For example, the map in Figure 6 below shows that the same ability (horizontal axis at the bottom of the map) to score 2 in Question 2 is required to score 1 in Question 1. Some questions (i.e. questions 3, 4, 5 and 8) are not mapped due to reversed thresholds (i.e. where a higher ability is required to score in a lower category somewhere on the scale). This usually occurs when marking guides or subjective interpretation thereof does not clearly distinguish what is required to score a particular score point rather than an adjacent score point. Such scores are often 'collapsed' for further analysis.

Figure 5: Category probability curves of Question 5

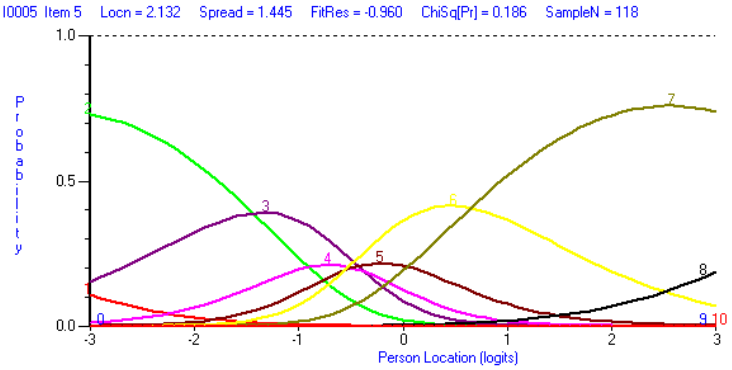
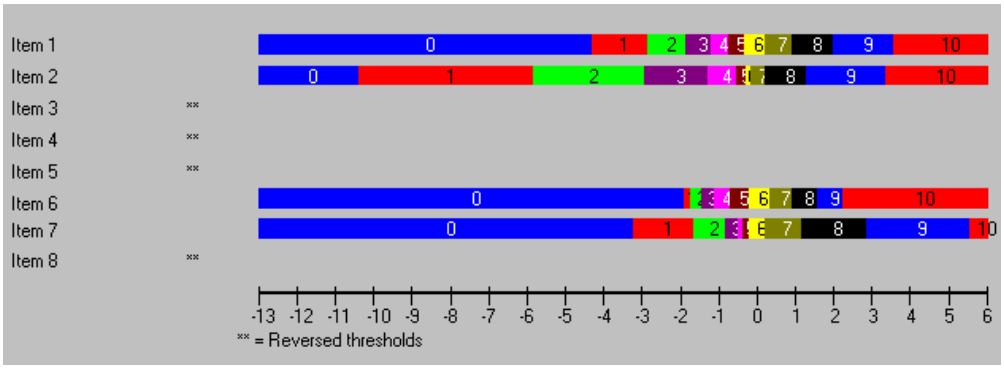



Figure 6: A threshold Map for the Eight Questions



These results highlight that the same category scores in different questions are not equivalent, especially at the lower and higher end of the scale.

Using the question difficulty estimates located on an interval scale through Rasch calibration, ability estimates for candidates can be obtained. Table 3 below summarises the ability estimates of the 11 candidates who had a total score of 21, together with the standard errors. To be able to relate the ability estimates to the CTT total scores, the ability estimates were transformed to z-scores and by using score equivalence tables, the z-scores were converted to 'Rasch-estimated' total scores shown in the bottom row of the table. The precision of the measures were comparable (see standard error values) and the abilities ranged from -0.51 logits to -0.20 logits (a difference of 0.31 logits). The Rasch-generated 'scores' ranged from 19.12 to 21.43 and since all candidates had total raw scores of 21, the differences in Rasch 'scores' can be ascribed to the combination of questions and their internal scoring categories that the candidates responded to.

Table 3: Ability estimates of the 11 candidates who scored 21 overall

| Candidate | One | Two | Three | Four | Five | Six | Seven | Eight | Nine | Ten | Eleven |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Ability** | -0.51 | -0.40 | -0.37 | -0.33 | -0.32 | -0.32 | -0.29 | -0.26 | -0.26 | -0.23 | -0.20 |
| **Std Error** | 0.35 | 0.33 | 0.34 | 0.35 | 0.36 | 0.36 | 0.31 | 0.37 | 0.34 | 0.34 | 0.33 |
| **Z-Score** | -0.25 | -0.10 | -0.07 | -0.02 | -0.01 | -0.01 | 0.03 | 0.07 | 0.07 | 0.11 | 0.14 |
| **Rasch** | 19.12 | 19.98 | 20.17 | 20.48 | 20.51 | 20.51 | 20.73 | 20.99 | 20.99 | 21.24 | 21.43 |

## Discussion

A standard method to obtain an overall score in an exam comprised of SAQs is to add the scores on the questions to obtain a total score for each candidate. Even when all the questions are compulsory, such methods are based on questionable assumptions. When scores are added at least interval level data is presumed (i.e. the relative value of each response category across all questions is treated as being the same and the unit increases across the rating scale is constant). This practice is further problematic if candidates have a choice of which questions they answer as adding scores on questions assumes that all questions are equal in difficulty, which is not the case as has been demonstrated in this study.

If the scoring structures within the questions are examined on an interval level scale, it is found that the difference in performance to score in an adjacent scoring category is not constant over all the scoring categories. For example, in Figure 6, it is noted that the difference between adjacent scores in the middle of the scale (i.e. between 3 and 7) is much smaller than the difference between adjacent scores on either end (i.e. from 0 to 2 and from 8 to 10). This suggests that since adjacent scores across individual questions require differing levels of ability, individual scores across questions cannot be compared directly.

It is also extremely unlikely that the same score category has the same meaning over questions. That is, a score of 6 in one question, for example, is unlikely to require exactly the same ability to score a 6 in all other questions. As such, not only do difficulties of individual questions need to be considered when determining an overall score for a candidate, the internal scoring category structure within each question also needs to be accounted for.

Through undertaking Rasch analysis, it is evident that the same total score of 21 derived from classical analysis for the 11 candidates is in fact, not an accurate representation of the candidates' true ability, since they selected different questions in the exam. Rasch calibration and analysis suggested for these candidates, the underlying 'true' scores ranged from 19.12 to 21.43, depending on the particular combination of questions answered. Taking the relative difficulties of the questions and the scoring structures into account with Rasch analysis, a difference of 2.31 in the overall score was found for the 11 candidates who all had total scores of 21, calculated from the initial classical analysis.

This has crucial implications for high stakes exams where a passing score can result in a candidate becoming certified to practice in a profession. If, for example, 21 was a passing score, then all 11 candidates would have passed based on their total scores derived from the classical analysis. When Rasch analysis was undertaken, only two (or four if 20.99 was considered as 21) would have passed.

This issue becomes more prominent if most scores are not clustered around the middle. Only 11 candidates with the same total scores were considered in this analysis but for larger samples, it would most likely increase the variance, making the differences in scores more apparent.

## Conclusion

This study highlighted that using Classical Test Theory to obtain a total score for examinees in order to compare performance is likely to advantage some candidates whilst disadvantaging others in the case where they had a choice of which questions they could answer. The same total score does not represent the same performance in such cases and Rasch analysis is more appropriate to identify the true total score of each candidate.

## References


Andrich, D. (1988). Rasch models for measurement. Newbury Park, CA: Sage.

Andrich, D. (1978). Rating formulation for ordered response categories. Psychometrika, 43(4), 561-573.

Barnard, J.J. (2012). A primer on measurement theory. Melbourne: Excel Psychological and Educational Consultancy.

Barnett-Foster, D., & Nagy, P. (1996). Undergraduate Student Response Strategies to Test Questions of Varying Format. Higher Education, 32(2), 177–198.

Bond, T.G., Fox, C.M. (2013). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed). New Jersey: Lawrence Erlbaum Associates.

Crocker, L.M., & Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: Holt, Rinehart and Winston Inc.

Hambleton, R.K., & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston: Kluwer-Nijhoff.

Hudson, R. D., & Treagust, D. F. (2013). Which form of assessment provides the best information about student performance in chemistry examinations? Research in Science & Technological Education, 31(1), 49-65.

Magno C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. International Journal of Educational Psychology Assessment [ED506058], 1(1), 1-11.

Mullen, K., & Schultz, M. (2012). Short answer versus multiple choice examination questions for first year chemistry. International Journal of Innovation in Science and Mathematics Education, 20(3), 1-18.

Petrillo, J., Cano, S.J., McLeod, L.D., & Coon, C.D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. Journal of the International Society for Pharmacoeconomics and Outcomes Research, 18(1), 25-34.

Tor, E., & Steketee, C. (2011). Rasch analysis on OSCE data: an illustrative example. AMJ 4(6), 339-345. http//dx.doi.org/10.4066/AMJ.2011.755.

Trigg, M., Barnard, J.J., Pham, H., & Devitt, P. (2016). Scoring short answer questions of five borderline

medical students. British Journal of Medicine and Medical Research,17(12), 1-7.
Ventouras, E., Triantis, D., Tsiakas, P., & Stergiopoulos, C. (2010). Comparison of Examination Methods Based on Multiple-Choice Questions and Constructed-Response Questions Using Personal Computers. Computers & Education, 54(2), 455-461.
Wainer, H., & Thissen, D. (1993). Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction. Applied Measurement In Education, 6(2), 103-18.
Wright, B.D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-166.

## The IEA: Researching education, improving learning

**Dr. Paulína Koršňáková**
Senior Research and Liaison Adviser to IEA

**Ms. Sive Finlay**
Communications Officer, IEA

The IEA, or the International Association for the Evaluation of Educational Achievement, is a nonprofit international scientific society that conducts comparative pedagogical research worldwide. Since 1958, IEA has measured students' achievement in subjects such as mathematics and science (TIMSS), reading (PIRLS), and civic and citizenship education (ICCS), investigated students' computer skills (SITES and ICILS), and researched early childhood (ECES) and teacher education (TEDS-M). The goal of our research is to gain a better understanding of education systems and to use this knowledge to help improve education policies and practices worldwide.

More than 60 countries are represented in the IEA network, and over 100 education systems participate in IEA studies. IEA studies are designed by educators for educators to answer questions such as: What do students know and what can they do? Is student achievement improving over time? What practices and policies are associated with student achievement? When a country participates, a representative sample of schools and students is invited to take the assessment. This approach allows IEA to assess educational progress within that country, working together with national coordinators to identify the challenges and opportunities, see what works, and to share practices and lessons from other participating education systems.

### What makes IEA different?

The IEA's mathematics, reading and science assessments are unique in the international study space because they are curriculum rather than age-based. This allows us to examine what students are expected to learn (intended curriculum), what is actually taught in schools (implemented curriculum), and student outcomes (achieved curriculum). IEA works with the national research coordinators of participating countries to ensure that what is tested is appropriate for their students. Assessment questions are pre-tested (referred to as pilot or field testing) before the assessment is administered to ensure high-quality comparative standards. IEA studies also explore school, home and other factors related to learning.

IEA studies are familiar to scholars and other education specialists within the subject areas of reading, mathematics and science, as well as experts interested in statistical analysis of large-scale data on educational achievement worldwide. Yet our studies do not have a high profile among researchers in other fields, many practitioners or the wider public. This is a pity, (and something we are actively trying to address) because IEA's published findings and open datasets provide a solid evidence base

**IEA Studies At A Glance**

| Study | Name | Assessment group | Frequency | More information |
|---|---|---|---|---|
| IEA TIMSS | Trends in International Mathematics and Science Study | 4th & 8th grade | Every 4 years, since 1995 | TIMSS & PIRLS International Study Centre, Lynch School of Education, Boston College |
| IEA PIRLS | Progress in International Reading Literacy Study | 4th grade | Every 5 years, since 2001 | TIMSS & PIRLS International Study Centre, Lynch School of Education, Boston College |
| ICCS IEA | International Civic and Citizenship Study | 8th grade | Two cycles completed (2009, 2016), currently recruiting for 2022 | ICCS website |
| IEA ICILS | International Computer and Information Literacy Study | 8th grade | One cycle completed (2013), currently in the main testing period for 2018 | ICILS website |
| IEA LaNA | Literacy and Numeracy Assessment | 4th grade (depends on the country) | On a project basis, LaNA is intended for developing educational systems as a stepping stone for participating in full IEA studies at a later date | IEA website |

for researchers, educators and policymakers interested in effecting change in their own countries.

### Applications of IEA's studies

Our mission is to provide educators, teachers and policymakers with insights into how students perform, and we are proud to work with partner organizations with similar goals. For example, results from our Progress in International Reading Literacy Study (PIRLS) 2016 were recognized by UNESCO as valuable insights for monitoring progress toward achieving Sustainable Development Goal (SDG) 4 which aims to "ensure inclusive and equitable quality education and promote

lifelong opportunities for all". These were summarized in an IEA and UNESCO joint publication, "Measuring SDG 4: How PIRLS can help".

PIRLS measures children's reading achievement at the end of grade four, which is close to the end of primary education in many countries. It is a time when students are transitioning from learning to read to reading to learn and therefore represents a key educational milestone. In addition to monitoring students' achievements, PIRLS also collects background information on how education systems provide educational opportunities to their students, as well as the factors that influence how students use these opportunities. This includes information about national curricula, students' home environment, socio economic status and teachers' education and training. These data provide valuable insights for monitoring progress in meeting the goals of SDG 4. For example, SDG target 4.2 states that, by 2030, all girls and boys should have access to quality early childhood development, care and pre-primary education so that they are ready for primary education. PIRLS 2016 results reveal that, on average, more children are attending pre-primary education and they tend to have higher reading scores than those who did not attend although this pattern is not the same across all countries. Download the full publication for more insights.

### Working with IEA's data

All of IEA's publications are open access and our data are freely available to anyone interested in educational research. The IEA Data Repository provides access to the data and accompanying documentation files of completed IEA studies. In addition, our International Database Analyzer (IDB) is a freely available software tool which enables users to work with data from most large-scale assessments, not just by the IEA but also those conducted by the US National Assessment of Educational Progress and the Organisation for Economic Co-operation and Development. The IDB Analyzer and accompanying video tutorials are available from the IEA website and the IEA also offers regular workshops on how to work with international large-scale assessment data.

### IEA's publications

IEA produces a range of free, open source publications, offering valuable reference tools and materials for researchers, practitioners and policymakers. These include the IERI Journal, *Large-scale Assessments in Education*, and the IEA Research for Education series of in-depth analyses based on IEA data, published in cooperation with Springer.

Interested authors should note that the publication costs for *Large-scale Assessments in Education* are covered by IERI so authors do not need to pay an article-processing charge. Full submission guidelines are available from the Springer website. Meanwhile, the strong download figures for the first few volumes of IEA Research for Education happily indicate that this series is already established as respected scholarly analyses to support informed policy advancement. A fourth volume will be released in May 2018.

In addition to our research publications, the IEA Compass Briefs in Education series publishes topical, accessible articles addressing key issues of interest to educational stakeholders, especially those involved in influencing educational decision and policymaking. Each publication in the series uses secondary analyses of IEA data to connect study results to recurrent and emerging questions in education at the international and national

levels. The IEA regularly publishes calls for proposals, and always welcomes thought-provoking concepts for the IEA Research for Education series or suggestions for topics for future Compass briefs. Access our latest publications here.

### Opportunities for researchers

The IEA is committed to encouraging and promoting high quality research based on IEA data. Our annual research awards recognize excellence in empirical research by graduate students, postgraduate students and established researchers working with IEA data. These awards were established in honor of Bruce H. Choppin and Richard M. Wolf to commemorate their significant contributions to IEA's mission. Applications are submitted by the end of March each year and authors of exceptionally innovative papers or theses may be invited to present their results at the prestigious IEA General Assembly meeting.

The IEA's International Research Conference is another valuable opportunity for researchers working with IEA data. Held every two years at different host institutions around the world, the conference offers a valuable forum for researchers to exchange ideas and information on critical educational issues in a comparative and global context. Our next conference will take place on 26-28 June 2019 at The Department of Education, Aarhus University in Copenhagen, Denmark. Proposals must be submitted by 30 September 2018 and conference registration will open in December, please visit the conference webpage for more details. We welcome applications from researchers, practitioners, policymakers or anyone who is interested in international large-scale assessments in education.

We look forward to seeing you in Denmark!

## Fifth in a Series - Interviews with ITC's Early Leaders:

Dr. Ronald K. Hambleton

*Editor's Note: Beginning in 2015, the ITC Council initiated an archives project to document the history of the organization. A series of questions were developed to elicit perspectives on the past, present, and future of the ITC, and these were sent to various individuals who have held positions of leadership over the years. In this issue of TI is the fourth in this series of interviews; Dr. Hambleton's responses to these questions were gathered in Spring 2018.*

*How did you become involved with the ITC and in what year, and what roles have you been active in over the years?*

I became involved with the ITC in 1982 and my close connection continued for 32 years. I was elected to serve on the ITC council several times with four year terms, and I was elected President for a four-year term (1990 to 1994) and with that election, I served four years as Vice President and four more years as Past President and I even served as Secretary for four years. I also served as IAAP's representative to the ITC Council. I served too as an early editor of our journal that appeared in the European Journal of Psychological Assessment. It all began for me when I was appointed APA's representative to the ITC. As a Canadian, working in the US,

with British parents and relatives, and teaching many international students, I was attracted to many international testing issues and practices. I think my work on the AERA, APA, and NCME Test Standards was the impetus for the APA appointment, and I was very happy to be nominated. Never did I expect 32 years later to be still involved.

*What were your initial impressions of the ITC, including its organizational structure, missions, and personnel?*

Well, in 1982, the situation with the ITC was that it was early in its development. There was a rough draft of a constitution and by-laws, an executive including Ken Miller, John Toplis, and Jac Zaal, and a small number of Council members including Elizabeth Nair from Singapore and Rolf Prinsloo from South Africa. Income was perhaps $2000 a year. As I was the APA representative, and because of APA's size and importance in the world of psychology, I was immediately appointed to the Council. I can say now that I was pretty disappointed with the situation. I judged the ITC to be without a clear set of goals, and no money to do much anyway. In my first report to the APA following the ITC Council meeting in 1982 I indicated that I thought the APA should seriously consider withdrawing their support. One activity where the ITC was successful in the early years was organizing symposia for meetings of the International Congress of Psychology and the International Association of Applied Psychology. These were fun to work on and somewhat useful to the testing field, but rarely did publications result, and so the impact of these efforts was small.

*Who were some of the key leaders in the ITC at that time and what were their roles?*

I greatly admired Ype Poortinga from the Netherlands but his direct involvement was minimal after 1982. He had been ITC President in the earlier years, and he had leadership skills and worldwide credibility as a famous cross cultural psychologist but he stayed on the margins of the organization by his own choice. Justin Schlegel from France edited the Bulletin for the ITC and did a wonderful job—He identified the issues for publication, edited the manuscripts, produced copies of the publication, and distributed them too. He was a workhorse with these publications but the overall impact was low because of the very small number of persons who received the publication. There was only limited effort on the part of Council members to expand the readership. Jac Zaal was an early contributor and later published a book with myself on current issues in the world of testing.

*What prominent changes have you seen in the ITC between when you first were a member and now?*

The leadership of the ITC through the executive and the Council has been outstanding and this has made all the difference since about 1980. I think of Tom Oakland, David Bartram, John Hattie, and Jose Muniz and others from the past who brought ideas, plans, and energy to the ITC. The 10 conferences the ITC has held have been more successful than we ever imagined, the website has been invaluable, and who would have predicted the great success of the conferences, the journal and newsletter publications, and the various sets of guidelines. Recent leadership from the Executive and Council has continued to be strong and is the best hope we have for the future of the ITC.

*Where has the ITC done well? What do you think the ITC's biggest accomplishment has been over the years?*

I have seen substantial growth in the ITC. Perhaps the biggest accomplishment has been the growth of an organization from very limited significance in the field of testing to what it has become today with conferences every two years bringing international leaders in the testing field together with scholars from all over the world to share ideas and advances, important and interesting publications such as Testing International and the International Journal of Testing, and the development and distribution of testing guidelines. For me personally, I am proud of our conferences, publications, and guidelines, but I too have appreciated the opportunity to develop my own skills particularly regarding testing around the globe. I am grateful too for the opportunity to develop friendships with so many colleagues—David Bartram, Jose Muniz, Tom Oakland, Jacques Gregoire, Fons von de Vijver, Ype Poortinga, and so many other outstanding scholars. To be working on behalf of the ITC with so many friends and colleagues in a collaborate way, made the past 32 years to be the best experience of my career.

*Where has the ITC possibly made "wrong steps" or mistakes?*

The ITC was just a few years old when I joined in 1982. Probably we made a number of mistakes. The failure to have a place for individual members limited the ways in which interested individual members could serve the ITC. I still think the membership is underutilized in the important work of the ITC. I had some strong negative feelings about the lack of productivity in the early years of the ITC. Progress was too slow and hampered by lack of vision and inactivity. Eventually, that situation changed in the 1980s and today I see the Council energized and moving in many directions and with nearly everyone contributing to the work. I will be watching with great interest, and saddened that my time is up.

*What do you perceive to be current challenges facing the ITC and what role should the ITC play in this regard?*

Testing is not having the impact that I believe it should have around the world. I'd like to see the ITC more involved in educating the leadership and the public in the many ways that improved testing could play in helping children, teachers, policy-makers, human resource specialists, and the public. I'd like to see closer relationships with PISA and TIMSS, for example, and other national and international assessment initiatives.
What do you believe the ITC should be focusing on right now and who are the key stakeholders that should be involved?
I think the ITC already has a number of important initiatives and I would like to see those continue. I would prefer to see guidelines development, conferences, and publications succeed and supported. I don't know if any major expansion is possible or desirable. Still, there is room for some minor initiatives. Representation by the ITC or their designees, at national meetings or regional meetings can have an impact on improving testing practices. Also, I would like to see the ITC capitalize on the skills of its members in ITC projects such as offering workshops at national and regional meetings, furthering the uses of testing guidelines, and testing projects, such as furthering testing with special populations.

*If there was one thing you could change with regards to the ITC, what would it be?*

I am not so familiar with ITC developments over the last four years, but I like everything that I see: excellent leadership from the Executive and Council, and lots of worthwhile activities.

*What structural or organizational changes to the ITC should be considered to further its effectiveness?*
I think the ITC organizational structure is fine, though I wish the Council could find some important ways to capitalize on the strength of its membership. I would like to see more involvement of the membership on task forces and committees to do the work of the ITC. Maybe we could establish regional centers of ITC activities.

*To what extent did you find it difficult to dedicate your time to the work of the ITC?*

When I joined the ITC in 1982 I was early in my career and besides that, not much was required each year of council members. Later as I become more involved, time was not much of a problem for me because the ITC had become an important part of my career.

*How would you compare the first ITC conference you ever attended with the more recent ones?*

I have attended all of the conferences so far, and the first one may have been the best one. The meeting was held at Oxford University in England and everyone wanted to be there. Oxford was a famous center for learning, and even if the conference had been a bust, it still would have been great to spend time at one of the most famous universities in the world. Also, everyone lived together in the dorms, and we all ate our meals together. You couldn't get away from interacting with colleagues from around the world, had you wanted to do that. The positive interactions and networking were an immensely important part of that first conference. Also, we had a great list of speakers—some of the best testing specialists in the world. We even published a book of proceedings. For many of us, that first conference was highly emotional and successful. But the conference in San Sabastian too was terrific. Paula Elosua provided a wonderful facility for the meeting, and she organized a terrific agenda. The social events were great too. I think the first and last conferences have been among the very best that we have ever had.
Among your various contributions, what do you believe may be your lasting legacy?
I hope my contributions overall have helped the ITC grow and become an important contributor to the improvement of testing practices around the world. I know I was viewed by some as a trouble-maker but I think my activities over the years proved to be useful and moved the ITC forward.

## About Ronald K. Hambleton

Ronald K. Hambleton holds the titles of Distinguished University Professor and Executive Director of the Center for Educational Assessment at the University of Massachusetts Amherst in the USA. He earned his B.A. degree (with Honors) in 1966 from the University of Waterloo in Canada with majors in mathematics and psychology, and an M.A. in 1967 and Ph.D. in 1969 from the University of Toronto with specialties in psychometric methods and statistics. He is a Fellow of Divisions 5 and 15 of the American Psychological Association (APA), a Fellow of the American Educational Research Association, and a Fellow of the International Association of Applied Psychology.

## Advancement of Psychometrics in Israel, with focus on the MOOC: "Development of Measurement and Assessment Tools"

**Avi Allalouf**
National Institute for Testing & Evaluation

### Background

In many countries around the world, apart from the US and perhaps a few other countries, the field of educational and psychological measurement (or "psychometrics") is not as developed as it should be in academia – The number of training programs is inadequate, and there are too few faculty members at institutions of higher education. In Israeli academe, there are also very few faculty members in this field, and there are no advanced degrees or training programs for the next generation of specialists. Those who want to study psychometrics generally do so in the US, and those who do not go to the US, have to rely on their place of work for training, usually in the skills and knowledge needed for a specific job.

In order to redress this situation, the Committee on Advanced Studies in Psychometrics in Israel, comprised of academics and psychometricians from various organizations, was set up. After discussing the situation and possible solutions, the committee had two main recommendations: (1) Providing scholarships in the field, or related fields, for doctoral and post-doctoral studies, in Israel and abroad, and (2) Developing a certification program in psychometrics in Israel. The committee suggested that the National Institute for Testing and Evaluation (NITE) would develop the certification program. The hope is that, within a few years, the cadre of those who have received scholarships and those who have completed the certification studies in psychometrics will strengthen and enhance this area in academic institutions and in professional organizations.

Fortunately, the financial backing needed to develop the certificate program is being donated by a philanthropic foundation approached by NITE. The foundation also provides academic scholarships in this field. Tuition fees (which are not high) will also be used toward covering the costs of developing and operating the certificate program.
The "Advancement of Psychometrics in Israel" is a project of NITE, It consists of two parts: scholarships and certified studies in psychometrics.

### Scholarships

The scholarships provide financial funding for one to four years. There are three tracks of scholarships:

- Scholarships for doctoral studies in Israel for three years. Three doctoral students accepted to this program in 2016 receive psychometric support for their thesis from experts in the field, and enroll in the certification studies in psychometrics, in addition to their doctoral studies;
- Scholarships for doctoral studies abroad for four years. One candidate who was accepted to the program in 2016 has started doctoral studies in the US; and;
- Scholarships for one year for post-doctoral studies abroad. Several candidates are being interviewed at the time of writing this

article. It is possible that additional scholarships will be offered at some future time.

*Certificate Studies in Psychometrics*

Certificate programs in psychometrics are intended for those working in the field of measurement and evaluation who are interested in professional development, as well as for those seeking to broaden their knowledge in these areas with the aim of entering the field in the future. These studies are not doctoral program. The study program is suitable for those working in the area of measurement and evaluation in the education system, in governmental ministries, public organizations and philanthropic foundations. The program is also intended for MA and PhD candidates interested in acquiring psychometric training for their research, and it is a requirement for doctoral students who have received a scholarship from the Committee for Advanced Studies in Psychometrics.

## Studies Objectives

The study program has been designed in such a way that those who complete it successfully will acquire theoretical, critical and practical knowledge and skills:

Theoretical: (1) knowledge of important terms in psychometrics as well as an understanding of their theoretical context and applications, (2) exposure to a wide variety of basic texts in psychometrics as well as new research directions in the field, and (3) awareness of the ethical issues in the field and their implications

Critical (1) independent and critical thinking skills in the area of measurement and evaluation, (2) ability to identify the ways in which testing affects – and is affected by – society

Practical: (1) familiarity with a wide variety of measurement and evaluation tools and with important principles in developing these tools, (2) statistical knowledge and skills needed for work and research in the field of measurement and evaluation, (3) enhanced expertise in the use of quantitative methods for research and problem solving, and (4) experience in writing, discussing and giving presentations about psychometric topics

## Structure of the Study Program

The study program is two years long. As a prerequisite, students must complete the following two courses (offered at several universities and colleges in Israel): "Introduction to Test Theory" and "Statistics and Advanced Research Methods".
The following five courses were developed especially for the certification study program:

1. The Foundations of Psychometrics: The course deals with basic topics in psychometrics – starting with interpreting correlation coefficients, reliability (including generalizability theory), and validity (including validity arguments and predictive validity studies). Later, the students become acquainted with scoring, equating and reporting of scores. The last part of the course focuses on factor analysis in the service of psychometrics.

2. IRT and Data Mining: In this course, students will be exposed to a selection of advanced data analysis methods used by psychometricians (mainly IRT and data mining). They will learn to understand the underlying theoretical principles of the different methods, to operate statistical programs for carrying out the analyses, and to properly interpret and present the results.
3. Practical Issues in Measurement and Evaluation: This course complements the first two courses offered. Its purpose is to provide

the knowledge and methodology needed for different aspects of psychometric research and practice, such as test administration, dealing with learning disabilities, preventing cheating, and differential item functioning. The course will also introduce students to the Standards for Educational & Psychological Testing (AERA, APA, & NCME, 2014).

4. The effect of testing on society: The goals of this course are to examine the many uses of tests and to engage in a critical discussion of their implications for society within historical, international, and Israeli contexts. The impact that educational and psychological tests have on society is looked at through the prism of other disciplines such as sociology, economics, communications, and law. The course deals with questions like: Should so much reliance be placed on tests? Do tests perpetuate gaps in society? Are tests fair? How will tests be viewed in the future?

5. "Development of Measurement and Assessment Tools" – a MOOC: This course is different from the others in the program. Its main objective is to teach participants how to create various assessment tools. The course is designed in such a way that it will be relevant and attractive to many target audiences that administer tests: school teachers, lecturers in academic institutions, as well as professionals from recruitment & selection companies and from public organizations. One of the main objectives of this initiative is to increase assessment literacy among the public. For this reason, much effort has been invested in developing this course as a MOOC – Massive Open Online Course. Many concerns related to the course were discussed: Which topics should be included? In what way can this MOOC supplement current textbooks on this topic How to approach heterogeneous populations? How can such a course be developed with a substantial but limited budget?


The course structure: The course comprises 50-videotaped teaching modules, each lasting about 15 minutes. A cadre of some 25 lecturers teach subjects in their area of expertise. Each module is accompanied by several questions that serve as formative assessment. In addition to the online material, there are five frontal sessions, involving mostly hands-on exercises in multiple-choice and open-ended item development, and in constructing assessments for non-cognitive skills. The course is divided into seven sections, each comprising several topics as seen below.

*Introduction, theory and practice*

- Introduction to the course; basic concepts
- Test types, reliability and validity
- Planning of test structure

*Cornerstones for building closed and open test items*

- Writing closed items and answer alternatives
- Writing open questions; preparing a scoring rubric
- Item analysis
- Building a complete test
- Tools and accommodations for those with learning disabilities

*Tests and tools in the education system*

- Tests and observation in school and classroom
- Standard tests in the education system
- International tests

*Assessment in the professional context; assessment centers & non-cognitive tests*

- Occupational analysis
- Assessment centers
- Scenario-based assessment using actors
- Interviews and biographical questionnaires
- Self-report personality tests
- Assessment of teamwork
- Assessment of colleagues
- Professional tendency questionnaires

*Intelligence and specific abilities*

- Assessment of writing by means of composition and closed test
- Assessment of second language proficiency
- Assessment of artistic skills
- Intelligence tests

*Attitude and values questionnaires*

- Theory and development of attitude questionnaires
- Theory and development of values questionnaires

*New areas of development*

- Computer-based assessment and computerized tests
- Computerized assessment for teaching and learning

## Status of the Certificate Studies and International Cooperation

Two of the courses (The Foundation of Psychometrics, and Development of Measurement and Assessment Tools) are currently taking place at NITE, and one course (Testing and Society) is currently taking place in a college. Starting from 2019, the materials of the course Development of Measurement and Assessment Tools will be available to everyone without tuition fees.

We believe that presenting the advancement of psychometrics project (scholarships and certification studies program) to the international community – via the readers of the ITC Newsletter "Testing International" – might enhance and expedite international efforts to develop and promote the field of psychometrics in other countries; it will hopefully also foster cooperation in the development of teaching tools for the field. The project has been presented in a European conference (AEA-E; Allalouf & Friedman, 2017). Most of the project material is in Hebrew, but we are thinking to translate some of the material to English, and to create a website. We look forward to sharing our experience with all.

I would like to thank Ruth Beyth Marom, Michal Beller, Tzur Karelitz – my colleagues on the Committee on Advanced Studies in Psychometrics in Israel, and Nitzan Friedman who works with me on the MOOC. I also thank Yoav Cohen, former general manager of NITE, and Anat Ben-Simon, the current general manager of NITE, for supporting the project.

## References

Allalouf, A., & Friedman, N. (November, 2017). Improving Assessment Literacy among Teachers and the General Public Using MOOCs, at the 17TH meeting of the Association of Educational Assessment – Europe (AEA-E), Prague, Czech Republic.

Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. Routledge.