AEA-Europe | Association for Educational Assessment - Europe

Social and Political underpinnings of educational assessment: Past, present and future

The 17th Annual AEA-Europe Conference

Programme | 2-5 November 2016

Limassol, Cyprus

aea EUROPE

University of Cyprus

AEA-Europe | Association for Educational Assessment – Europe
www.aea-europe.net

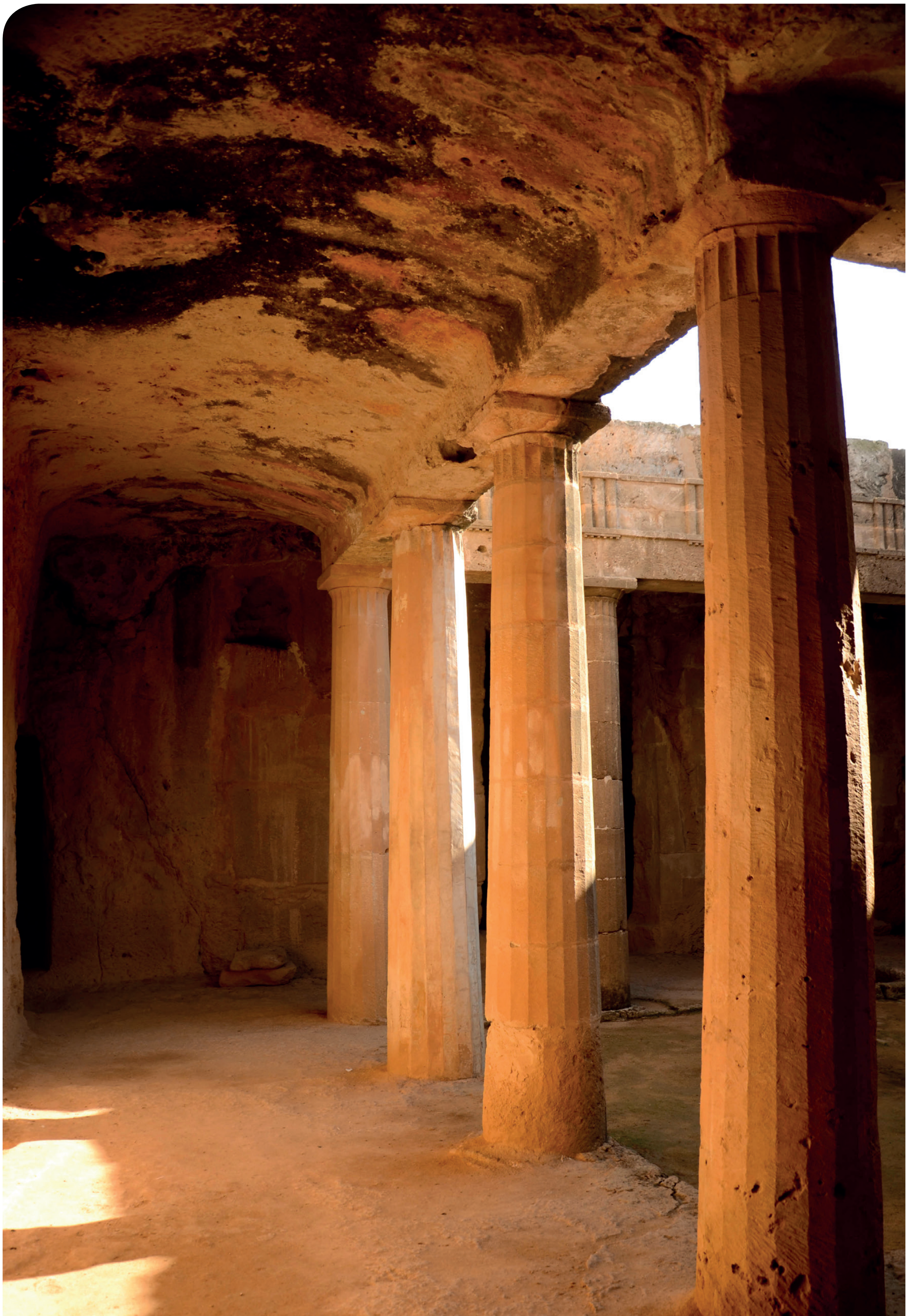
President | Guri Nortvedt
Department of Teacher Education and School Research University of Oslo, Norway

Vice President | Thierry Rocher
Directorate for Assessment, Forecasting and Performance (DEPP), France

Executive Secretary | Alex Scharaschkin
AQA, United Kingdom

Treasurer | Cor Sluijter
Cito, The Netherlands

Contents



Introduction

Social and Political underpinnings of educational assessment: Past, present and future

This year's AEA-Europe conference will provide opportunities for advancing ideas from across Europe and beyond related to the social and political underpinnings of educational assessment. We are meeting in Limassol, Cyprus. Limassol is an ancient town, the third largest in Cyprus. This town is home to more than a hundred educational institution. Why are we here?
The topic of our conference is how social needs and agents as well as political needs and ideas influence assessment. The citizens of Limassol have stood strong under the influence of many external forces. In medieval time, the crusaders took a seat and were followed by the venetians and the ottomans. Establishing an educational system under foreign reign must have been a challenge, with the strong social and political forces present in these circumstances. A Greek school was established in Limassol in 1819, followed by the first public school in 1841 and a girls' school in 1861. Today the people of Cyprus are among the highest educated in Europe. And, as in all European countries educational assessment is under discussion, being influenced by social forces, lobby groups, university admission policies, parents, politicians, international comparative assessments. The list of participants is long, but this is an important debate and participation, openness and awareness is vital. Our theme 'Social and Political underpinnings of educational assessment: Past, present and future' is a tribute to the history of education and educational assessment in Limassol, in Cyprus and to our host the University of Cyprus.

Educational assessment policies and programmes develop within particular historical, political, economic and social contexts. The wide range of national and international educational assessments must therefore be viewed as complex socio-political phenomena. To analyse and comprehend how assessment programmes and policies evolve, it is important to reflect on how each assessment is shaped and transformed by social and political agents interacting at the national and international level.

Policy makers and politicians are often influenced by powerful social forces, driven by ideological debates, social theories, and market and industry requirements. In addition, stakeholders such as the testing industry and teachers' unions do what they can to influence decision making and educational policies. Historical and pragmatic considerations, such as the need for functional and easy-to-explain high-stakes assessments to allocate scarce resources, often influence how and on what students are assessed. Consequently, assessment is often used as an instrument for public management policy, framed within a general rhetoric of the need for accountability and transparency. However, the consequences of assessment systems need to be carefully considered as they can otherwise be dire.
For instance, high-stakes assessments might be transformed into a de facto filtering system that prevents social mobility and facilitates social stratification.

Depending on the use of the assessment outcome, the many different types of educational assessment (including classroom assessment, national exams, university entrance examinations, tests for certification and licensure, and international comparative tests) can be high- or low-stakes for individuals and society. In the last decade, the assessment community has been provided with new technical tools and measurement models, enabling for instance new assessment modes. However, rapid innovations in information technology and psychometrics are not always coupled with solid theoretical advancements.

Consequently, the validity of assessments might be threatened as more attention is often given to the financial and political aspects of the development rather than to potential unintended consequences of introducing new or reshaping existing assessment programmes. In addition, financial, political and ethical debates about privacy and the use of big data for assessment purposes albeit necessary, can be challenging when so many stakeholders are involved. For instance, it has been observed that technological and psychometric innovations are particularly difficult to communicate to the wider public, not only to parents and students but also to politicians and sometimes even to the research community. Moreover, in order to satisfy diverse political agendas, we have repeatedly seen an over-confidence regarding the possibility of reliably measuring complex abilities, skills and attitudes. This over-confidence occasionally leads to public disappointment, media pressure and distrust, and heated political debates about ‘dropping standards’ and ‘failing assessment systems’.

At the level of every-day classroom assessment, the last decades have nurtured very promising assessment initiatives such as Assessment for Learning, and more countries are moving toward a greater emphasis on assessment made by the teachers themselves. Coupled with innovative software and elaborate feedback techniques, emphasis is now given to the professional development of teachers, classroom assessment and student learning. However, we are still struggling to accumulate a critical mass of empirical evidence to show that all this investment has led to societies with more equal opportunities and social justice.

Each country in Europe offers a different environment within which issues of educational assessment can be viewed. At the 2016 AEA-Europe conference in Limassol participants will present and reflect on research on assessment and its relationship with policy and practice in many different contexts, for instance how countries tackle the educational implications of issues of equity, ethnicity, access, cultural awareness, or development of new assessment formats. Given that national and international educational assessment systems operate within specific legal and cultural frameworks that are the products of ever-changing social and political processes, the assessment community and policy makers share responsibility and should all be held accountable for the consequences of educational assessment. As an assessment community, we have an ethical responsibility to strive for development of more effective and socially fair assessment practices; however, we will only be able to do this if we understand the nature and functions of the social and political contexts that are the main drivers of our work.

Following our tradition, the conference programme comprises keynotes, open paper sessions, discussion groups and poster presentations in addition to pre-conference workshops. This year the programme also includes two mini symposia. The topic of social and political underpinnings of assessment has attracted a very large number of proposals that has surprised all involved in organising the conference. Each year, more proposals are made, illustrating the actuality of our conference themes. Consequently, it has been necessary to extend the programme and include also a parallel session on Saturday morning. We take this as an indication of the importance of our theme, and welcome the many opportunities such a rich programme will provide. Unfortunately, a rich programme also forces every participant to carefully choose which opportunities to take and what discussions to enter into. Still, many possibilities to share knowledge, experiences and views with colleagues from Europe and beyond will exist.

The council of AEA Europe are grateful for the contributions of our organising committee, the local organising committee, the Scientific Programme Committee, EasyConferences and our sponsors. You will find the names of the people involved in securing a good conference at the back of this conference book and the logos of our main sponsors on the back cover. Without them, we would not have succeeded in making this conference.

Guri A. Nortvedt

Guri A. Nortvedt
President, AEA-Europe

Programme

Wednesday 2nd November

9.00 – 9.30 **Coffee and registration**
Registration (St Raphael reception area)

9.30 – 16.30 The pre-conference workshops take place in St Raphael Hotel

Workshop 1: **Applications of Item Response Theory *(Atrium A)***

Presenters: *Theo Eggen, Frans Kleintjes and Marieke van Onna (Cito, The Netherlands)*
The workshop will offer an introduction to IRT and applications from a practical point of view. IRT is used for many measurement applications including item banking, test construction, adaptive test administration, scaling, linking and equating, standard setting, test scoring and score reporting. Main features of these applications will be addressed in the workshop. Participants will be able to understand and assess the usefulness of IRT in their own work.
The workshop is aimed at those who want to know more about IRT with a focus on applications. Participants might be novice ore more experienced user. No prior knowledge is required to attend the workshop. Participants will practice using the software for some examples and are invited to bring their own laptops for practicing (Windows).

Workshop 2: **Developing Constructed Response Test Items *(Atrium B)***

Presenter: *Ezekiel Sweiry (AQA, United Kingdom)*
While considerable research and guidance on writing selected response (SR) test items exists in the literature, guidance on writing constructed response (CR) items is scarce. This is despite the fact that there is far greater potential for examinees to misunderstand the requirements of CR items. In addition, the development of CR mark schemes that show both intrinsic validity and high levels of marker agreement presents serious challenges that are all but absent for SR items.
The purpose of this workshop is to present and discuss guidance on developing CR items and their mark schemes.
The workshop is aimed at anyone with an interest in constructed response item and mark scheme design, including test developers, educational assessment researchers and those involved in the scoring of responses. No specific prior knowledge is needed.

Workshop 3: **Discussing and sharing experiences of the social, political and also cultural drivers of assessment practices and policies *(Megaron B)***

Presenters: *Egil Hartberg, Vegard Meland (Lillehammer University College, Norway) and Stephen Dobson (University of South Australia, Australia)*
Workshop participants will have the opportunity to discuss and share their varied experiences of working with different assessment resources. This workshop is for Educational professionals working in the tertiary and schooling sectors with responsibility/experience of assessment. The workshop will also be relevant for assessment developers in the corporate and state sector who would like to

discuss experiences of building educator capacity through innovative assessment resources and how this might impact upon reducing social inequality amongst stakeholder groups.

Workshop 4: **Using data collected in IEA studies for informing policy and practice *(Megaron G)***

Presenters: *Sabine Meinck (IEA Data Processing and Research Center, Germany) and David Rutkowski (University of Oslo, Norway)*
The workshop will illustrate possibilities and limitations of large-scale assessment data to inform policy and practice, given the complexities of the designs of such studies. Participants will get the opportunity to learn about the study specifics and develop and exchange ideas on how results arising from this data can be 'translated' best to inform politicians, school staff and the public. This workshop is for researchers and education specialists using ILSA data for informing policy and practice are addressed in this workshop.

18.30 – 19.30 **Registration *(St Raphael reception area)***

18.30 – 19.00 **Welcome reception for new attendees *(St Raphael Resort, hotel lobby and outside veranda*)**

19.00 – 20.00 **Welcome reception *(St Raphael Resort, hotel lobby and outside veranda)***

Thursday 3rd November

8.30 – 9.15 Coffee *(lobby)*

8.30 – 9.30 Registration *(lobby)*

9.30 – 10.10 Welcome addresses *(Panorama)*

- *Professor Costantinos Christofides, Rector, University of Cyprus (Cyprus)*
- *Professor Andreas Demetriou, former Minister of Education and Culture of Cyprus (Cyprus)*
- *Guri A. Nortvedt, AEA-Europe President (Norway)*

10.10 – 11.05 Keynote presentation *(Panorama)*

Chair: Guri A. Nortvedt (Norway)
Title: Innovative assessment approaches and new paradigms within the assessment of teachers' professional competencies
Professor Gabriele Kaiser, University of Hamburg (Germany)

11.05 – 11.30 Coffee *(lobby)*

11.30 – 12.15 Keynote presentation *(Panorama)*

Chair: Iasonas Lamprianou (Cyprus)
Title: New Developments and Techniques in Educational Assessment
Professor George Marcoulides, University of California, Santa Barbara (USA)

12.30 – 13.30 Lunch

13.30 – 15.00 Open paper sessions A, B, C, D and E, symposium 1 and discussion group 1

Session A: Impact of Assessment Design on Equity and Social Justice *(Atrium A)*
Chair: Theo Eggen (The Netherlands)

1. Should assessment be an agent of social change?
   Gordon Stobart (Institute of Education, United Kingdom)
2. Do tiered examinations affect candidates' achievement? Some empirical evidence on Modern Foreign Languages
   Nadir Zanini (Cambridge Assessment, United Kingdom)
3. The fairness of our tests: reporting on item-level DIF analysis of high-stakes exams in England
   Ben Smith and Ruth Johnson (AQA, United Kingdom)

Session B: Characteristics of Questions/Items *(Atrium C)*

Chair: Cor Sluijter (The Netherlands)

4. Question quality: The concept of quality in the context of exam questions
   Victoria Crisp, Martin Johnson and Filio Constantinou (Cambridge Assessment, United Kingdom)
5. How do question writers compose examination questions? Question writing as a socio-cognitive process
   Martin Johnson, Filio Constantinou and Victoria Crisp (Cambridge Assessment, United Kingdom)
6. Investigating experts' perceptions of examination question demand
   Tom Bramley (Cambridge Assessment, United Kingdom)

Session C: Improving standards *(Megaron A)*

Chair: Dina Tsagari (Cyprus)

7. Applying formal concept analysis in assessment: can it help mediate between socio-political and technical understandings of the meaning of exam grades?
   Alex Scharaschkin (AQA, United Kingdom)
8. Improving the maintenance of standards in England: evaluating a comparative judgement approach to awarding
   Kate Kelly, Charlotte Stephenson, Neil Stringer and Faith Jones (AQA, United Kingdom)
9. Assessing Primary Writing in a Politically High Stakes Context
   Sarah Maughan (AlphaPlus, United Kingdom)

Session D: Supporting teacher's and rater's assessment practices *(Megaron B)*

Chair: Guri A. Nortvedt (Norway)

10. Exploring Teachers' Approaches to Classroom Assessment: An Instrument Development Study
    Christopher DeLuca, Danielle Lapointe-McEwan, Adelina Valiquette and Andrew Coombs (Queen's University, Canada)
11. Supporting Teachers' Assessment Professionalism in the Context of Change in Curriculum and Examinations
    Kay Livingston, Louise Hayward, George MacBride, Carolyn Hutchinson and Ernest Spencer (University of Glasgow, United Kingdom)
12. Using comparative judgement for the assessment of academic writing: an examination of its validity
    Tine van Daal, Marije Lesterhuis, Liesje Coertjens, Vincent Donche and Sven De Maeyer (University of Antwerp, Belgium)

Session E: Conceptions of subject difficulty and subject/test taking strategies *(Megaron G)*

Chair: Rolf Vegar Olsen (Norway)

13. Inter-subject comparability: How does adjusting grade boundaries affect schools, subjects and candidates in England?
    Caroline Lau, Simon Eason, Ben Jones (AQA, United Kingdom) and Mike Cresswell (Independent, United Kingdom)
14. Subject Entry Choices and Perceptions of Subject Difficulty: Are the Two Linked, and if so, How?
    Benjamin Cuff (Ofqual, United Kingdom)
15. Test Taking Practices, Background Variables, and their Relationship to Validity
    Elena Papanastasiou and Agni Stylianou-Georgiou (University of Nicosia, Cyprus)

Symposium 1

Error in high-stakes assessments *(Phoenix)*

Chairs: Isabel Nisbet, Paul Newton, Beth Black, Sarah Hughes and Stuart Shaw (United Kingdom)

1a. Talking about assessment error
Isabel Nisbet (University of Cambridge, United Kingdom)

1b. Tolerating difference of opinion
Beth Black and Paul Newton (Ofqual, United Kingdom)

1c. To Review or to Remark – that is the question: detecting error when reviewing students' marks
Sarah Hughes and Stuart Shaw (Cambridge Assessment, United Kingdom)

Discussion group 1

Educational assessment for the future: Asia-Pacific and Scandinavian contexts *(Atrium B)*

Tony Burner (University of South East Norway, Norway), Nhat Ho Thi and Duyen Tran (Hanoi National University of Education, Vietnam)

15.00 – 15.30 Coffee break *(lobby)*

15.30 – 17.00 Open paper sessions F, G, H, I, J and K and discussion group 2

Session F: Impact of educational policy on equity and social justice *(Atrium A)*

Chair: Paul Newton (United Kingdom)

16. An exception that proves (tests) the rule: social justice and national standardised assessment policy in Scotland
Louise Hayward, Kay Livingston, George MacBride and Ernest Spencer (University of Glasgow, United Kingdom)

17. School characteristics, SES and achievement
Trude Nilsen (University of Oslo, Norway) and Jan-Eric Gustafsson (University of Gothenburg, Sweden)

18. Teacher quality mediating and moderating the relation between SES and achievement in Nordic countries
Hege Kaarstein, Trude Nilsen (University of Oslo, Norway) and Jan-Eric Gustafsson (University of Gothenburg, Sweden)

Session G: Measurement of complex skills *(Atrium C)*

Chair: Andrew Boyle (United Kingdom)

19. The Primary Scientific Reasoning Test – In Pursuit of Content Validity
Diana Ng (Oxford University Centre for Educational Assessment, United Kingdom)

20. Multistage testing and Disability Act: a new test method for policy evaluation
Reinaldo Dos Santos and Thierry Rocher (French Ministry of Education, France)

21. The use of images in rating scales to assess attitudes, feelings and dispositions
Christine Merrell (CEM, University of Durham, United Kingdom) and Peter Tymms (University of Durham, United Kingdom)

Session H: Changes in assessment systems *(Megaron A)*

Chair: Bas Hemker (The Netherlands)

22. Recovery from reform: The 'Sawtooth Effect' in UK secondary school assessments
Michelle Meadows and Benjamin Cuff (Ofqual, United Kingdom)

23. High stakes testing and social responsibility: an investigation into the balance of high and low order skills in high-stakes tests in England and Wales
Martin Walker (CEM, University of Durham, United Kingdom) and Kate Crabtree (Qualifications Wales, United Kingdom)
24. Social and political underpinnings of educational assessment: Admission to medical schools
Avital Moshinsky and Naomi Gafni (NITE, Israel)

Session I: Assessment – reforms and innovations *(Megaron B)*
Chair: Steven Bakker (The Netherlands)

25. Quality evidence of initial teacher education programmes: Aligning standards and graduate teachers' experiences in ever-changing social and political arenas
Claire Wyatt-Smith and Anna Du Plessis (Learning Sciences Institute Australia, Australia)
26. Preparing for College Success: Exploring the Impact of the High School Cambridge Acceleration Programme on US University Students
Magda Werno and Stuart Shaw (Cambridge Assessment, United Kingdom)
27. Snapshots of deep learning over time: A novel approach to measuring student progress
Ian Jones (Loughborough University, United Kingdom) and Brian Henderson (No More Marking Ltd., United Kingdom)

Session J: The social effects of assessment results *(Megaron G)*
Chair: Gill Stewart (United Kingdom)

28. Unleashing the power of human capital: Workforce assessment, strategy, skills and standards
Daniela Muresan (ETS Global, The Netherlands)
29. Supporting stakeholder trust in A level Modern Foreign Language outcomes: is there a native speaker effect?
Rachel Taylor (Ofqual, United Kingdom)
30. PISA Results as a Mirror of Social Stratification Reproduction: A Story from Serbia
Jelena Radišić (University of Oslo, Norway), Aleksandar Baucal (University of Belgrade, Serbia) and Jasminka Čekić Marković (Center for Education Policy, Serbia)

Session K: The micro-politics of assessment *(Atrium B)*
Chair: Christina Wikstrom (Sweden)

31. The complex interplays between assessment and learning that shape writing development during the transition to university from A-level
Natalie Usher (OUCEA, United Kingdom)
32. The micro-politics of the school in the process of changing assessment cultures
María Teresa Flórez Petour (University of Chile, Chile)
33. Monitoring of student achievements in Mathematics as an effective instrument to adjust individual learning paths for students and to enhance didactical tools of teachers
Laila Issayeva, Daniyar Temirtassov, Baimurat Akhmetov, Yerbol Nurguzhin (Nazarbayev Intellectual Schools, Kazakhstan), Nico Dieteren and Frans Kleintjes (Cito, The Netherlands)

Discussion group 2

National Improvement Framework and Assessment of Children's Progress *(Phoenix)*
Kit Wyeth, Jane Gallacher, Donna Bell (Scottish Government, United Kingdom) and Graeme Logan (Education Scotland, United Kingdom)

17.15 – 19.00 **Meet and greet for doctoral students *(Megaron B and foyer)***
Host: Stéphanie Berger (Switzerland) and the Professional Development Committee

Friday 4th November

8.30 – 9.00 Coffee *(Lobby area)*

9.00 – 10.30 **Open paper sessions L, M, N, O and P and discussion group 3**

Session L: Assessment quality *(Atrium A)*
Chair: Stephen Dobson (Australia)

34. Improving marking quality and examiner experience: zoning scripts with generic answer booklets
Matthew Glanville (International Baccalaureate, United Kingdom) and Martin Adams (RM Results, United Kingdom)
35. What makes a good seeding script? Perceptions from Principal Examiners of an UK awarding body
Martina Kuvalja and Simon Child (Cambridge Assessment, United Kingdom)
36. Implementing the Assessment Agenda for Intermediate Vocational Education in the Netherlands: Developing Flexible Digital Exams
Cor Sluijter and Marieke van Onna (Cito, The Netherlands)

Session M: Assessment policies: what's behind? *(Megaron B)*
Chair: Stéphanie Berger (Switzerland)

37. Examining the 'global revolution' of English language learning
Elizabeth Shepherd, Wahida Amin and Victoria Ainsworth (British Council, United Kingdom)
38. Educational governance and PISA: challenges and prospects in the Cypriot context
Myria Vassiliou, Aristotelis Zmas (European University of Cyprus, Cyprus) and Michalis Michaelides (University of Cyprus, Cyprus)
39. The process of social construction of assessment policy: the case of SIMCE
María Teresa Flórez Petour, Jenny Assael Budnik and Cristian Cabalín Quijada (University of Chile, Chile)

Session N: Applying assessment data to inform teaching *(Megaron A)*
Chair: Yoav Cohen (Israel)

40. Diagnostic information from national assessment: Exploration of a simple cognitive diagnostic model
Daniël Van Nijlen and Rianne Janssen (KU Leuven, Belgium)
41. Programme for International Student Assessment (PISA): Results and cha(lle)nges for the Cyprus educational system
Salome Hadjineophytou (Saint Louis University, Cyprus)
42. MathemaTIC – a multilingual, digital and adaptive environment for increasing equity in mathematics learning in Grades 5 and 6 students in Luxembourg
Philippe Arzoumanian (DEPP, Ministry of Education, France) and Amina Kafaï-Afif (Agency for Development of Quality in Schools, Ministry of Education, Luxembourg)

Session O: Assessing what matters – validity challenges in assessment *(Megaron G)*

Chair: Frans Kleintjes (The Netherlands)

43. A new approach to the reform of vocational qualifications in Wales by Europe's newest qualification regulator
Cassy Taylor (Qualifications Wales, United Kingdom)
44. How valid are pre-university vocational qualifications?
Rose Clesham (Pearson, United Kingdom)
45. Digital responsibility – a required literacy for citizenship: How to understand and measure the concept?
Ove Edvard Hatlevik and Inger Throndsen (University of Oslo, Norway)

Session P: Assessment and learning in the digital age *(Atrium B)*

Chair: Isabel Nisbet (United Kingdom)

46. Measuring the Impact on Learners: evaluating a whole school reading programme in the UK
Grace Grima, Elpida Ahtariou, Krystina Dunn, Vanessa Greene (Pearson UK, United Kingdom), Sue Bodman, Glen Franklin, Jane Hurry and Catherine Carroll (UCL, United Kingdom)
47. Exploring the effects of undertaking the Extended Project Qualification
Charlotte Stephenson (AQA, United Kingdom)
48. Investigating student-student interactions in an assessment of collaborative problem solving: An in-depth analysis of think-aloud protocols
Ronny Scherer and Fazilat Siddiq (University of Oslo, Norway)

Discussion group 3

Standard-setting/maintaining and public trust in national examinations around the world *(Phoenix)*

Lena Gray (Centre for Education Research and Practice, AQA, United Kingdom), Tina Isaacs (UCL Institute of Education, United Kingdom), Jo-Anne Baird (University of Oxford, United Kingdom), Dennis Opposs (Ofqual, United Kingdom), Christina Wikstrom (Umeå University, Sweden) and Anton Beguin (Cito, The Netherlands)

Meet the publication committee *(Atrium C)*

Chair: Gill Stewart (United Kingdom)

10.30 – 11.45 **Poster presentations and coffee *(Panorama)***

Chair: Cor Sluijter (The Netherlands)

A. Lower secondary school teachers' and students' conceptions of assessment purposes
Georgia Solomonidou (Independent, Cyprus) and Michalis Michaelides (University of Cyprus, Cyprus)
B. Does the mode of standardisation matter? The effect on reliability of marking and marker perceptions
Lorna Stabler, Magda Werno, Sarah Hughes and Stuart Shaw (Cambridge International Examinations, United Kingdom)
C. The impact of lack of legislation on educational assessment: a case study
Joaquin Cruz (University of Jaen, Spain)
D. A Contribution of Case Study Tests to Predicting Performance on Real-Life English Language Tasks
Elena Sokolova, Elena Iureva and Oksana Shahhoud (Russian State Social University, Russia)

E. A review study to identify adaptive algorithms for increasing the efficiency of Comparative Judgement
San Verhavert, Vincent Donche, Sven De Maeyer and Liesje Coertjens (University of Antwerp, Belgium)

F. Developing a framework for user generated assessment
Mark Frazer and Sarah Gott (CEM, University of Durham, United Kingdom)

G. The effect of teaching referencing practices on student attitudes towards cheating
Rebecca Hamer and Tamsin Burbidge (International Baccalaureate, The Netherlands)

H. Investigating students' Perception on Rubric-oriented Assessment in the Micro Context of Russian University
Olga Mironova (Nizhny Novgorod Linguistics University, Russia)

I. How is ICAEW, a global chartered accountancy body, replicating the realities of the workplace in professional exams without compromising quality and rigour?
Mike Green (RM Results, United Kingdom)

J. Initiative of criteria-based assessment system implementation in Kazakhstan
Olga Mozhayeva, Aidana Shilibekova and Aliya Mustafina (AEO 'Nazarbayev Intellectual Schools', Kazakhstan)

K. You read on screen, I read on paper – Are we reading the same texts?
Ragnhild Engdal Jensen (University of Oslo, Norway)

L. A site of tension: the complex case of GCSE English speaking and listening
Ruth Johnson (AQA, United Kingdom)

M. The Internationalization of Higher Education: assessing university staff
Victoria Levchenko (Samara University, Russia)

N. Provision of feedback in L2 exam classes in Cyprus
Dina Tsagari and George Michaeloudes (University of Cyprus, Cyprus)

O. Computerized standard setting using the Data-Driven Direct Consensus (3DC)
Jesse Koops, Remco Feskens and Frans Kleintjes (Cito, The Netherlands)

P. What make PISA items more difficult for students with minority background? Analysing the effects of item interactivity and response format in a computer-based assessment of scientific literacy
Nani Teig (University of Oslo, Norway)

11.45 – 12.45 General assembly *(Phoenix)*

12.45 – 13.45 Lunch

13.45 – 15.15 Open paper sessions Q, R, S, T and U, symposium 2 and discussion group 4

Session Q: Good and bad consequences of assessment *(Atrium A)*
Chair: Gordon Stobart (United Kingdom)

49. Teacher evaluation – trapped between accountability and learning: Assessing teacher professionalism – formatively
Sølvi Lillejord, Kristin Børte (Knowledge Centre for Education, Norway) and Therese Hopfenbeck (Oxford University Centre for Educational Assessment, United Kingdom)

50. Understanding and Developing Relational Aspects of Assessment for Policy and Practice
Ruth Dann and Jo Basford (Manchester Metropolitan University, United Kingdom)

51. The teacher as a stakeholder in utilizing an assessment for learning tool
Guri A. Nortvedt and Anubha Rohatgi (University of Oslo, Norway)

Session R: Social and political context of assessment *(Atrium C)*

Chair: Thierry Rocher (France)

52. Social, political and cultural impact of high-stakes national assessments: the case of UNT in Kazakhstan
    Aigul Yessengaliyeva (PhD Sociology, Kazakhstan) and Nico Dieteren (Cito, The Netherlands)
53. Four Issues in the Debate about Admissions Testing and the Four P's: Psychometrics, Politics, the Press and the Public
    Yoav Cohen and Anat Ben-Simon (NITE, Israel)
54. Student Conceptions of Understanding and of Assessment Supporting Learning for Understanding
    Rebecca Hamer (International Baccalaureate, The Netherlands) and Erik Jan van Rossum (Independent researcher, The Netherlands)

Session S: Assessment policy reforms *(Megaron A)*

Chair: Louise Hayward (United Kingdom)

55. Changes in school accountability measures – Is there an effect on Enquiries about Results behaviour?
    Sara Humphries, Vikas Dhawan and Beth Black (Ofqual, United Kingdom)
56. Assessment and policy discord: transition to secondary level education in Northern Ireland
    Leanne Henderson (Queen's University, United Kingdom)
57. Reforming educational assessment in Trinidad and Tobago
    Bas T. Hemker, Cor Sluijter (Cito, The Netherlands) and Newman Burdett (Independent, United Kingdom)

Session T: Talking about the fundamental issues when validating assessments *(Megaron B)*

Chair: Stuart Shaw (United Kingdom)

58. Validity arguments for considering different expectations when setting cut-off scores in a formative assessment in digital responsibility
    Ingrid Radtke (Vox – Norwegian agency for lifelong learning, Norway) and Ove Hatlevik (University of Oslo, Norway)
59. Validation of student selection system in Kazakhstan: transparency and accountability
    Miras Baimyrza, Zamira Rakhymbayeva (Nazarbayev Intellectual Schools, Kazakhstan), Caroline Jongkamp and Frans Kleintjes (Cito, The Netherlands)
60. Helping the industry to engage with validity and validation
    Paul Newton (Ofqual, United Kingdom)

Session U: Supporting teachers in assessment *(Megaron G)*
Chair: Jannette Elwood (United Kingdom)

61. Computer-based formative assessment in classrooms
Stéphanie Berger (University of Zurich, Institute for Educational Evaluation/ University of Twente, Research Center for Examination and Certification, Switzerland), Urs Moser (University of Zurich, Institute for Educational Evaluation, Switzerland), Angela Verschoor (Cito, The Netherlands) and Theo J.H.M. Eggen (Cito, The Netherlands/University of Twente, Research Center for Examination and Certification, The Netherlands)
62. Our TALE: the importance of the social and educational context of assessment
Dina Tsagari (University of Cyprus, Cyprus), Karin Vogt (University of Heidelberg, Germany), Ildiko Csepes (University of Debrecen, Hungary), Tony Green (University of Bedfordshire, United Kingdom) and Nicos Sifakis (Hellenic Open University, Greece)
63. Integration of standards for expected writing proficiency within an AfL-approach
Ragnar Thygesen (University of Agder, Norway), Lars S. Evensen and Gustaf B. U. Skar (Norwegian University of Science and Technology, Norway)

Symposium 2

Social and political underpinnings of admissions procedures to higher education – the perspectives of five European countries *(Atrium B)*
Chairs: Giray Berberoğlu (Turkey), David Gabelaia (Georgia), Iasonas Lamprianou (Cyprus), Christina Wikstrom, Per-Erik Lyrén, Magnus Wikstrom (Sweden), Avi Allalouf and Naomi Gafni (NITE, Israel)

2a. Major Problems, Policy Implications and Possible Solutions in the University Admission System in Turkey
Giray Berberoğlu (Başkent University, Turkey)
2b. University Entrance Examinations in Georgia
David Gabelaia (National Examinations and Assessment Center (NAEC), Georgia)
2c. University entrance examinations in Cyprus: a battlefield for lobbies and political proxy wars?
Iasonas Lamprianou (University of Cyprus, Cyprus)
2d. Meritocracy vs egalitarianism: validity challenges in the selection to higher education in Sweden
Christina Wikstrom, Per-Erik Lyrén and Magnus Wikstrom (Umeå University, Sweden)
2e. The Effect of Politics on Higher Education Admissions – Israel
Avi Allalouf and Naomi Gafni (NITE, Israel)

Discussion group 4

Developing Teacher Assessment Capacity: Diverse Perspectives from around the World *(Phoenix*)
Lisbeth Brevik (University of Oslo, Norway), Christopher DeLuca (Queen's University, Canada), Christine Harrison (King's College, United Kingdom), Carolyn Hutchinson, Kay Livingston (University of Glasgow, United Kingdom), Sandra Johnson (Assessment Europe, United Kingdom) and Claire Wyatt-Smith (Australian Catholic University, Australia)

15.15 – 15.45 Coffee break (lobby)

15.45 – 17.15 Open paper sessions V, W, X, Y and Z and discussion group 5

Session V: Scales, scores, analysis *(Atrium A)*
Chair: Michalis Michaelides (Cyprus)

64. *TIA-Excel – an easy tool for test and item analysis in classroom assessments*
    Eef Ameel and Rianne Janssen (KU Leuven, Belgium)
65. Elimination scoring as an alternative for correction for guessing in multiple-choice questions: an empirical comparison
    Rianne Janssen, Jef Vanderoost and Tinne De Laet (KU Leuven, Belgium)
66. Number identification: A scale identifying a progression pathway
    Sarah Gott, Lee Copping, Helen Cramman, Christine Merrell (CEM, University of Durham, United Kingdom) and Peter Tymms (University of Durham, United Kingdom)

Session W: Methodological advances in assessment *(Megaron A)*
Chair: Sarah Maughan (United Kingdom)

67. High-stakes assessment instruments: one-size-fits-all versus flexibility
    Caroline Jongkamp and Angela Verschoor (Cito, The Netherlands)
68. Innovations in standardized testing in Lithuania. Measuring Higher order Thinking Skills in Lithuania
    Eglé Melnike (NEC, National Examinations Centre Lithuania, Lithuania) and Frans Kleintjes (Cito, The Netherlands)
69. Predicting item difficulty: methodological challenges and way forward
    Yasmine El Masri, Jo-Anne Baird (Oxford University Centre for Educational Assessment, United Kingdom), Steve Ferrara (Pearson USA, USA) and Peter W. Foltz (University of Colorado, USA)

Session X: Accountability in assessment *(Megaron B)*
Chair: Michelle Meadows (United Kingdom)

70. Shifting Emphases: qualifications, accountability and school improvement
    Tim Oates and Sylvia Green (Cambridge Assessment, United Kingdom)
71. 'If you join them, you don't have to beat them'
    Birgitte Arctander Stub, Mari Bjugstad Wiken and Ida Large (Norwegian Directorate for Education and Training, Norway)
72. Improving students' future prospects or extending the reach of the accountability framework? Investigating the impact of the English Baccalaureate on the educational landscape
    Emma Armitage (AQA, United Kingdom)

Session Y: Addressing equal opportunities regarding participation in assessment situations *(Megaron G)*
Chair: Therese Hopfenbeck (United Kingdom)

73. Providing Cross-national Comparability of Test Results for International Assessment in Higher Education
    Elena Kardanova, Irina Brun, Denis Federiakin (Higher School of Economics, Russia) and Prashant Loyalka (Stanford University, USA)
74. Patterns in the use of languages for differentiated learning of mathematics by primary school students in Luxembourg
    Catalina Lomos (Luxembourg Institute of Socio-Economic Research (LISER), Luxembourg) and Amina Kafaï-Afif (Agency for Development of Quality in Schools, Ministry of Education, Luxembourg)
75. Assessing individual participation in collaborative group work
    Ayesha Ahmed (University of Cambridge, United Kingdom) and Ruth Johnson (AQA, United Kingdom)

Session Z: Comparative judgment *(Atrium B)*

Chair: Tom Bramley (United Kingdom)

76. A methodology for applying students' interactive task performance scores from a multimedia-based performance assessment in a Bayesian Network
Sebastiaan de Klerk (eX:plain/University of Twente, The Netherlands), Bernard Veldkamp (University of Twente, The Netherlands) and Theo Eggen (Cito/University of Twente, The Netherlands)
77. D-optimal adaptive comparative judgement
Yaw Bimpeh (AQA, United Kingdom)
78. Comparative Judgement and Scale Separation reliability: Yes, but what does it mean?
San Verhavert, Sven De Maeyer, Vincent Donche and Liesje Coertjens (University of Antwerp, Belgium)

Discussion group 5

The quality of school-based vocational assessment *(Phoenix)*

Andrew Boyle (AlphaPlus Consultancy Ltd., United Kingdom), James Morgan (Scottish Qualifications Authority, United Kingdom), Bernadette Dilger (Universität St. Gallen, Germany) and Jean-Pierre Jeantheau (Agence Nationale de Lutte Contre l'Illettrisme, France)

19.30 – 23.00 Conference dinner *(Traditional Restaurant)*

(busses will depart from St Raphael at 19.00)

Saturday 5th November

8.30 – 9.00 Coffee *(lobby)*

9.00 – 10.30 Open paper sessions AA, BB, CC, DD and EE and discussion group 6

Session AA: Assessment accountability *(Atrium A)*
Chair: Grace Grima (United Kingdom)

79. Cross-validating teachers' judgments and test-based results
Carolyn Hutchinson (University of Glasgow, United Kingdom) and Sandra Johnson (Assessment Europe, France)
80. The transition in England from high stakes to low stakes assessment in primary school science
Oliver Stacey (NFER, United Kingdom)
81. Policymakers Intentions for Test-Based Accountability Policy – the Test as an 'Omni-Instrument'
Lisa Amdur (Tel Aviv University, Israel) and Irit Mero-Jaffe (Beit Berl Academic College, Israel)

Session BB: The moderation of teacher assessment *(Megaron A)*
Chair: Theo Eggen (The Netherlands)

82. Improving moderation of teacher assessed work
Chris Wheadon (No More Marking Ltd, United Kingdom) and Daisy Christodoulou (Ark Schools, United Kingdom)
83. Thinking outside the box: Using e-marking to moderate internally assessed coursework through dynamic sampling
Matthew Glanville (International Baccalaureate, United Kingdom) and Thomas Kelly (RM Results, United Kingdom)
84. Evidence for the reliability of coursework
Tom Benton (Cambridge Assessment, United Kingdom)

Session CC: Assessment – social implications *(Megaron B)*
Chair: Tim Oates (United Kingdom)

85. Developing an assessment for 4 year-olds – challenges and tensions
Heather Bamforth and Catherine Kirkup (NFER, United Kingdom)
86. Reconsidering young people as social and political agents in educational assessment reform: Students' voices in national assessment transformations
Jannette Elwood (Queen's University, United Kingdom)
87. Randomised Controlled Trials: how assessment research can contribute to evidence-based policy and social justice
Andrew Boyle (AlphaPlus Consultancy Ltd., United Kingdom)

Session DD: Technical, political and social stakes when transferring to e-assessment *(Megaron G)*
Chair: Amina Kafaï-Afif (Luxemburg)

88. Utilising technology in the assessment of collaboration: A critique of PISA's collaborative problem solving tasks
Simon Child and Stuart Shaw (Cambridge Assessment, United Kingdom)
89. Political stakes during the transition towards computer-based assessments: the case study of a large-scale online assessment of 160,000 students in France
Sandra Andreu and Thierry Rocher (DEPP, France)

90. Learning in Digital Networks – A novel assessment of students' ICT literacy
Fazilat Siddiq (University of Oslo, Norway) and Perman Gochyyev (University of California, Berkeley, USA)

Session EE: Background factors influencing student achievement and the success of schooling *(Atrium B)*
Chair: Peter Tymms (United Kingdom)

91. Academic achievement and subjective well-being in primary school children
Tatjana Kanonire (Institute of Education, National Research University, Higher School of Economics, Russia)

92. The progress of first-year school-children: looking for factors of educational inequalities in the beginning of schooling
Alina Ivanova, Inna Antipkina and Elena Kardanova (National Research University Higher School of Economics, Russia)

93. Development of reading skills in pre-school students: the role of parental investments
Marina Vasilyeva (Boston College, USA), Alina Ivanova and Elena Kardanova (National Research University Higher School of Economics, Russia)

Discussion group 6

Are the influences from social and political agents beneficial and a necessity in the development and validation of educational assessments? *(Phoenix)*
Anna Lind Pantzare and Pia Almarlind (Umeå University, Sweden)

10.30 – 11.00 Coffee *(Lobby area)*

11.00 – 11.45 **Keynote presentation *(Panorama)***
Chair: Alex Scharaschkin (United Kingdom)
Title: Multimedia-based Performance Assessment in Dutch Vocational Education
Dr. Sebastiaan de Klerk, eX:plain (The Netherlands)

11.45 – 12.00 Break

12.00 – 12.45 **Keynote presentation *(Panorama)***
Chair: Thierry Rocher (France)
Title: Collaboration and Agency in Educational Assessment
Gudrun Erickson, University of Gothenburg (Sweden)

12.45 – 13.30 **Awards and closing session *(Panorama)***
Guri A. Nortvedt (Norway) and Thierry Rocher (France)

13.30 – 14.15 Lunch

Keynotes | Abstracts and Biographies

Thursday 3rd November

10.15 – 11.00 Innovative assessment approaches and new paradigms within the assessment of teachers' professional competencies
Professor Gabriele Kaiser (University of Hamburg, Germany)

Recent research on the professional competencies of mathematics teachers, which has been carried out during the last decade, is characterized by different theoretical paradigms on the conceptualization and assessment of teachers' professional competencies, namely cognitive versus situated approaches. Building on the international IEA Teacher Education and Development Study in Mathematics (TEDS-M) and its Follow-up study, TEDS-FU, the presentation will describe cognitive and situated approaches to assessing the professional competencies of teachers. In TEDS-FU the cognitive oriented framework of TEDS-M has been enriched by a situated orientation including the novice-expert framework and the noticing concept as theoretical approaches on teachers' competences for analysing classroom situations. Correspondingly, the assessment instruments were extended by using video-vignettes for assessing teachers' perception, interpretation and decision-making competencies in addition to cognitive oriented knowledge tests. The presentation will discuss the different kinds of theoretical paradigms, namely cognitively oriented versus situated oriented. Based on this distinction the necessity to develop new innovative assessment formats will be argued for, namely video-based assessment instruments getting closer to actual teaching situations and thus evaluating situated facets of the professional competencies of teachers complementing common knowledge tests with open and closed items focusing on knowledge-based facets of the professional competencies of teachers.
The staged video-vignettes focus on these situated competence facets evaluating the teachers noticing competences, i.e. to perceive particular events in an instructional setting, interpret the perceived activities in an instructional setting to develop decision options either as anticipating a response to students' activities or as proposing alternative instructional strategies.
The strengths and weaknesses of both assessment approaches will be contrasted based on the instruments and the results of TEDS-M and TEDS-FU.

Furthermore, connecting the results of TEDS-FU with TEDS-M allows comprehensive insight into the structure and development of the professional competencies of mathematics teachers, the complex interplay between the different facets of teachers' competencies and the high relevance of teaching practice for the development of these competencies. The presentation will show on the one hand that both approaches – cognitive and situated – are needed for a comprehensive description of teachers' professional competencies. On the other hand it will be shown that both approaches can be integrated in a prolific way.

The prospects will discuss possible cultural elements of these assessment paradigms drawing on newly established studies broadening TEDS-FU towards the evaluation of structural relations between teachers competencies and students achievement gains mediated by instructional quality (TEDS-INSTRUCT). First tentative results of a national study transferring TEDS-INSTRUCT to another Federal State of Germany (TEDS-VALIDATE) will be described as well as first insight from an international study aiming to transfer these studies into East Asia (TEDS-EAST-ASIA).

Biography
Gabriele Kaiser holds a master's degree as a teacher for mathematics and humanities for lower and upper secondary level, which she completed at the University of Kassel in 1978 with the first state degree. After having worked at school and completion of the second state degree, she worked as a scientific assistant at the Department of Mathematics at the University of Kassel, where she completed her doctorate in mathematics education (rer. nat.) in 1986 with a study on applications and modelling supervised by Werner Blum and Arnold Kirsch. Based on a grant for Postdoctoral Research by the German Research Society (DFG) she undertook her post-doctoral

study in pedagogy on international comparative studies at the University of Kassel, which she completed in 1997. From 1996-1998 she hold a guest professorship at the University of Potsdam. Since 1998, she is full professor for mathematics education at the Faculty of Education of the University of Hamburg.

Her areas of research include modelling and applications in school, international comparative studies, gender and cultural aspects in mathematics education and empirical research on teacher education. Gabriele Kaiser's most recent projects deal with teacher education, partly under an international perspective. Together with Sigrid Blömeke and Rainer Lehman (both Humboldt-University Berlin) she has carried out from 2006-2010 the IEA Teacher Education Study in Mathematics (TEDS-M), which compares the efficiency of teacher education in various countries. This project was supported by the German Research Society (DFG). She was already participating from 2004-2008 in the Pilot Study for the TEDS-Study, which was developing first instruments for this ambitious international study, the so-called Mathematics in the 21st Century study. Related to this study qualitatively oriented supplementary studies on future teachers' professional knowledge and their comparison have been extended to Hong Kong and Australia in collaboration with Gloria Stillman and Jill Brown from Australia, Ngai-Ying Wong and Issic Leung from Hong Kong. Most recently Gabriele Kaiser has carried out a follow-up-study to the German part of the study TEDS-M together with Sigrid Blömeke, Johannes König and Martina Döhrmann, in which the cohort of future teachers tested in TEDS-M is followed into their practical work in school, a project funded by the German Research Society as well. This project (TEDS-FU) used an enriched framework of TEDS-M and additional video-based evaluation instruments. Furthermore, she was participating in another study led by Sigrid Blömeke, in which the framework of TEDS-M was extended to future teachers of mathematics, German and English (TEDS-LT), measuring the professional knowledge of future teachers in the transition from Bachelor to Master studies (funded by the German Ministry for Education and Research, BMBF). A study funded by the German Telekom Foundation examined the development of the professional knowledge of future mathematics teachers at the beginning of their study in the frame of innovative teacher education projects. This study was carried out jointly with Sigrid Blömeke, Rainer Lehmann and Hans-Dieter Rinkens from 2008 to 2012. Her most recent research focus on the structural relation of the professional competencies of practicing teachers measured with instruments developed in the study TEDS-FU and the gain of students' achievements taught by the evaluated teachers. The assessment of the instructional quality of the teachers in focus is carried out by classroom observations and serves as mediator variable. This study is currently carried cooperatively with Sigrid Blömeke and Johannes König out in the Federal State of Hamburg, named TEDS-INSTRUCT. A validation study of this design aiming to evaluate the validity of the framework and the instruments in the Federal State of Thuringia has just been implemented, carried out jointly with Sigrid Blömeke and Johannes König and funded by the German Ministry of Education and Research (TEDS-VALIDATE). A supplementary study funded by the European Union in the frame of the Marie Skłodowska-Curie programme evaluates, whether the theoretical framework and the instruments developed in Western countries can be transferred to East Asian countries and whether the structural relations hold for Chinese teachers and their students taught too (responsible researcher: Xinrong Yang) (study EAST-WEST).

In October 2010 she took up the position as Vice Dean of the Faculty of Education being responsible for research, promotion of young researchers and international cooperation. Since 2005 she serves as Editor-in-chief of ZDM Mathematics Education (formerly Zentralblatt fuer Didaktik der Mathematik), published by Springer. She is Convenor of the 13th International Congress on Mathematics Education (ICME-13), which will take place 2016 at the University of Hamburg expecting several thousands of mathematics educators.

11.30 – 12.15 New Developments and Techniques in Educational Assessment

Professor George A. Marcoulides (University of California, Santa Barbara, USA)

New developments and techniques in the field of educational assessment continue to propagate at an incredible rate. Computer information technology has played a key role in changing the overall educational assessment landscape. This technology is used to administer different types and formats of assessments, for the tailoring of assessments to individual examinees, for the automated scoring of items, for the automated scoring of essays and constructed responses, for the automated evaluation of speech, for the automated development, assembly and banking of new items, and for the development of new algorithms and models to be used in these various assessment applications. Indeed, it is hard to imagine anyone these days who has not had to deal with at least one of these new technological innovations. Some scholars have described these technological innovations as 'providing the greatest promise for furthering ...knowledge', and as 'perhaps the most important and influential... revolution to have occurred...' Unfortunately, the connection between many of these innovations and practice is not always apparent. This talk will provide an overview of selected educational assessment advances and highlight the link between technological innovation and practice. The talk will also discuss various methodological misunderstandings concerning these innovations that clearly warrant careful consideration in terms of their potential political, economic, and even social ramifications.

Biography

George A. Marcoulides is Distinguished Professor of Research Methodology in the Department of Education at The Gervitz Graduate School of Education, and a member of the Quantitative Methodology in Social Sciences Program at the University of California, Santa Barbara. He was previously a Professor of Research Methods and Statistics at the University of California, Riverside and a Professor of Statistics in the Department of Information Systems and Decision Sciences at California State University, Fullerton (CSUF). He has also been a visiting professor at the University of California-Los Angeles, the University of California-Irvine, the University of Geneva, and the University of London. He has served as a consultant to numerous educational authorities, government agencies, companies in the United States and abroad, and to various national and multi-national corporations. He has co-authored or co-edited 15 books and edited volumes, published over 200 articles in scholarly journals and books, and presented numerous times at national and international conferences. His contributions have received Best Paper Awards from the Academy of Management, the Decision Sciences Institute, and the American Educational Research Association–University Council for Educational Administration. He is a Fellow of the American Educational Research Association, a Fellow of the Royal Statistical Society, and a member of the Society of Multivariate Experimental Psychology. He is currently Editor of the journals Structural Equation Modeling and Educational and Psychological Measurements, Editor of the Quantitative Methodology Book Series, and on the editorial board of numerous other scholarly journals.

Saturday 5th November

11.00 – 11.45 Multimedia-based Performance Assessment in Dutch Vocational Education

Dr. Sebastiaan de Klerk (eX:plain, The Netherlands)

During this keynote, I will present the results of my four-year PhD research on the introduction of a new type of assessment, which we have called Multimedia-based Performance Assessment (MBPA), in Dutch vocational education and training (VET). An MBPA can be considered as a computer-based performance assessment. The assessment incorporates multiple forms of multimedia and can be used to measure skills that would normally be measured in a traditional performance-based assessment (PBA). PBA is a very common form of assessment in Dutch VET. The goal of the presented research was to investigate whether the MBPA could be a more efficient and effective way of assessing performance skills. Both qualitative and quantitative evidence is presented to substantiate the use of MBPA in Dutch VET. Qualitative research consists of literature study on the measurement properties of PBA, the current state of MBPA and simulation-based assessment, and the design and development op MBPA. Quantitative research consists of two experimental studies in which a sample of students have both performed in a PBA and two MBPA's. First, a validity study has been carried out to investigate to what extent the MBPA can be used to measure skills that are required to perform a specific vocation (i.e., confined space guard) and that are currently being measured in a PBA. Finally, in a psychometric study, using Bayesian network modeling, the measurement properties of the MBPA and its latent ability structure have been determined.

Biography
Sebastiaan de Klerk currently works as a researcher at eX:plain, a foundation that broadly aims to improve measurement practice in vocational education, and the Research Center for Examinations and Certification, a collaboration between the University of Twente and Cito (the Dutch national institute for measurement). His research mainly focuses on the development and evaluation of simulation-based assessment in the vocational education domain. Sebastiaan obtained his PhD in educational measurement (January 2016) from the University of Twente in the Netherlands.

12.00 – 12.45 Collaboration and Agency in Educational Assessment
Professor Gudrun Erickson (University of Gothenburg, Sweden)

My presentation focuses on the dual function of educational assessment, to enhance and improve learning – for students as well as for teachers – and to contribute to fairness and quality. In spite of fundamental similarities, these functions are often treated as two completely separate phenomena, frequently characterized as formative and/or summative assessment, or assessment for learning and/or of learning, respectively. In my presentation I will talk about the balancing act between the two, emphasizing common principles that need to guide the development of a wide range of practices of assessment with the purpose of enhancing learning as well as equity in a broad sense – at individual, pedagogical and societal levels. In this, the concept of reciprocity will be brought into the discussion. I will base my reasoning on five fundamental questions that, from a general perspective, help guide the planning and analysis of all types of assessment, namely Why?, What?, How?, Who? & And...?. Thus, I will be talking about aspects of aims, constructs, methods, agents and uses, including consequences, of assessment. My main focus will be on issues of collaboration and agency, in particular on the role of students, in continuous assessment as well as in the development of large-scale summative materials. In this, I will touch upon aspects of power and empowerment, politics and policy, object and subject, connecting my reasoning to validity and ethics in a broad sense. The empirical basis for my presentation is, on the one hand, experiences from the development of large-scale assessment materials of different kinds within the Swedish national system, on the other hand, examples from European projects, where students' opinions have been collected, analysed and utilized. In this, I will also make some reference to international examples of assessment materials where student agency is in focus, as well as to guidelines for good practice in educational testing and assessment.

Biography
Gudrun Erickson is Professor of Education in Language and Assessment at the Department of Education and Special Education, University of Gothenburg, Sweden. Originally a teacher of languages, with long experience of teaching and teacher education, as well as national curriculum development. Commissioned by the Swedish National Agency for Education as project leader for the Swedish national testing and assessment programme for languages, comprising materials for formative as well as mandatory, summative use. GE has been, and is currently, engaged in a number of European projects focusing on learning, teaching and assessment. Between 2013 and 2016, President of the European Association for Language

Testing and Assessment (EALTA). Gudrun's main research interest is collaborative approaches to the development of language assessment and testing procedures and materials, in particular focused on issues of agency and contributions by test-takers.

Open papers | Abstracts

Open papers | Abstracts

Thursday 3rd November

Session A: Impact of Assessment Design on Equity and Social Justice

1. **Should assessment be an agent of social change?**
Gordon Stobart (Institute of Education, United Kingdom)

The intention of the paper is to generate discussion about the historical and current role of assessment in supporting and/or challenging social stratification.

The paper briefly examines the concept of social stratification and the forces that maintain it, drawing on Bourdieu's (1979) ideas of social and cultural capital and how they relate to assessment practices. Two examples are used to explore the issues. The first is the role of high-stakes selection testing either perpetuating social stratification or challenging it. The claim is that this form of assessment has a mixed record. A case study of Chinese civil service selection examinations, dating back over a thousand years, is used as an example of how an assessment system can undermine social stratification by reaching out beyond elite groups.

Current high-stakes assessments are then discussed in relation to their fairness to different social groups and how results are interpreted in terms of individual merit, with no recognition given to differences in resources and preparation. While a greater diversity of groups can now take examinations does social stratification continue to be reflected in the curriculum (what is studied), in marking and grading (what is rewarded)? What of the privileged preparation of some groups (tutoring/'crammers') and in the way results are interpreted?

The second example looks at the history and use of IQ testing in supporting social stratification. From Binet's 0riginal pragmatic and diagnostic approach to improve learning, intelligence testing was rapidly introduced into the Anglo-Saxon world to reinforce social beliefs that some groups, the rich and privileged, were dominant because of superior intelligence. The corollary of this was that the poor, and other racial groupings, had limited intelligence which led to their lower socio-economic status. To compound this many of the leading IQ testers were advocates of eugenics (a term coined by Galton) and campaigned to restrict reproduction of the poor and 'feeble-minded' as well as opposing immigration of certain ethnic groups. Policies based on socially stratified differences in IQ are still advocated.

These examples are the background to a discussion of the role of current assessment policy and practice in relation to social stratification and change. When, and to what extent, should assessment be used as a agent of change? Where are there examples of assessments reducing social stratification and providing access for groups who may previously been disadvantaged? Is there a role of positive discrimination in areas such as university entrance or job recruitment? How should IQ and ability testing be viewed in relation to social stratification?
The final section considers ways in which assessments can minimise the effects of social stratification. This involves making tests fairer for candidates from different social groupings. This begins with access questions: who gets taught and by whom? Curricular questions ask about whose knowledge is taught and in what way are other cultures treated (Apple, 1989). Assessment questions consider the form of the assessments and what is rewarded.
This scrutiny leads to Joy Cumming's questions: When setting standards and test content, are we really sure this is the knowledge that we need?

Are we really privileging certain knowledges to maintain a dominant culture, and in doing so, ensuring perpetration of ourselves, as people who have succeeded in the formal educational culture to date? (2000, p. 4)
The positive message of the paper is that, while we may never achieve fair assessment, we can make it fairer in relation to students from different social groupings.

References

- Bourdieu, P. (1979) Distinction: A Social Critique of the Judgement of Taste. Translated 1984. Harvard: Harvard University Press.
- Cumming, J. (2000) ‘After DIF, What Culture Remains?’ 26th IAEA Conference, Jerusalem.

2. Do tiered examinations affect candidates’ achievement? Some empirical evidence on Modern Foreign Languages

Nadir Zanini (Cambridge Assessment, United Kingdom)

The General Certificate of Secondary Education (GCSE) sat by 16 year olds in England is usually taken by students from a wide ability range (Bishop et al, 1999). Most GCSEs use differentiated papers targeted at different abilities within a two-tiered model; with the Foundation tier spanning grades C-G and the Higher tier spanning grades A*-D.

The use of tiering in GCSE assessments, however, has not been free from criticism. In addition to comparability issues associated with awarding overlapping grades at Foundation and Higher tier (Good & Cresswell, 1988; Wheadon & Beguin, 2010; Bramley, 2014), the need for teachers to select the tier of entry raises the possibility that candidates may be inappropriately entered for the Foundation tier, which could damage students’ aspirations and prevent them from achieving the best grade they are capable of (Baird et al., 2001). Considering that some universities and Further Education colleges require a B grade (which is not available to Foundation tier candidates) at GCSE for progression, the use of tiering in GCSE assessments could hinder efforts to ensure equality of opportunity in education and limit social mobility. This has recently led to a public debate on alternative models for differentiation, involving, among others, the Department for Education and the Office of Qualifications and Examinations Regulation (Ofqual). Notwithstanding, in the newly reformed GCSEs, tiering will remain.

This study investigates how tiering is currently functioning in GCSE Modern Foreign Languages (French, German and Spanish), specifications which are due to be reformed. In the new MFL assessments, the weighting of tiered components on the total score will be increased. More specifically, the aim of this study was threefold. First, to investigate whether, and to what extent, there is evidence of students’ achievement being capped due to the measurement characteristics of the current MFL tiered assessments. Secondly, to study the determinants of the tier of entry, considering a broader set of covariates than those used in previous studies

(Gillborn & Youdell, 2000; Wilson & Dhawan, 2014). Thirdly, to check whether there is an 'entry to Foundation tier' effect on achievement, once other variables connected to tier of entry are accounted for.

For each subject in turn, three strands of analysis were conducted on data from the June 2014 examination session: i) the computation of a 'notional' B grade boundary to identify how many candidates entered at Foundation level might have in fact been able to achieve a B grade; ii) multi-level logistic regression to study the determinants of the tier of entry; iii) propensity-score matching to retrieve the 'entry to Foundation tier' effect.

For the three subjects considered, the key findings of this work are highlighted below:

- Overall, the tiering structure of the current GCSE MFL specification did not affect candidates' achievement. For the three subjects considered, it emerged that only a very small number of candidates who were entered at Foundation tier would have achieved more than a C grade had they been entered at Higher tier.
- The study of the link between tier of entry and candidates' background characteristics showed that gender, as well as prior and concurrent attainment, were key factors affecting tier of entry, with more able female candidates less likely to be entered for the Foundation tier.
- The comparison of Foundation and Higher tier candidates' achievement on un-tiered units highlighted that, while differences reduced noticeably when background characteristics had been accounted for, they did not disappear completely. This suggests the presence of a small but significant effect of the 'entry to Foundation tier' that may be partly due to students' exposure to specific teaching material.

3. The fairness of our tests: reporting on item-level DIF analysis of high-stakes exams in England

Ben Smith and Ruth Johnson (AQA, United Kingdom)

Within the current policy context in England, there is a move away from assessments which combine examinations with school-based assessment, to assessments which rely wholly on externally marked, terminal examinations. The rationale for this policy shift has included a range of arguments, but one significant claim that has been made is that external examinations are intrinsically fairer (DfE, 2012); research evidence suggests however that this is not automatically the case. Previous studies have instead found that certain types of assessment items result in different outcomes for different social groups (Solando-Flores and Turnbull, 2008; Twist and Sainsbury, 2009). If some marks are less accessible to certain candidates based solely on their underlying demographic characteristics then the validity of an assessment is called into question. If this is currently the case within exams in England, a policy shift that places an increases weighting on external exams, rather than on internal assessment (which can be more readily adapted to different cultural contexts) has the potential to increase rather than reduce unfairness.

This paper reports on a research project that contributes to this body of research through examining two high-stakes assessments at item level to identify items, or types of items, which appear to advantage or disadvantage groups of students. The two assessments chosen as the site for this research are AQA GCSE English and Maths, selected because of the particularly high stakes nature of these two assessments. Achievement at a grade C or higher in GCSE in Maths and English is frequently demanded (alongside other measures) by gatekeepers to further and higher education and to employment. Achievement in English and Maths also dominates governmental measurement of schools through performance tables (DfE, 2014).

The effect of a range of contextual and demographic factors (including gender, ethnic group and socio-economic status) on item responses was analysed using a logistic regression framework. In order to facilitate a more nuanced understanding of differential item functioning (DIF), latent class logistic regression methodology was also utilised (Zumbo et al, 2015), which allows for the possibility that not every member of a certain social group will exhibit the specific traits that make them more or less likely to perform well on a given assessment. This methodology is somewhat similar to exploratory factor analyses, in that latent classes (subgroups) of candidates for whom an item shows variations in DIF can emerge from the analysis of items – something traditional logistic regression methodology will not capture. Deeper conclusions can thus be drawn about which social groups might be affected by any differential validity within an item by examining the characteristics of candidates in each latent class (though it is worth noting that latent classes do not emerge for all items: in many, logistic regression methodology is sufficient to explain any DIF).

Items and mark schemes in the assessments considered were also coded using a number of parallel typologies which considered the different dimensions of the items. On one level the codes used bluntly categorise items (for example by length of response or number of marks), and on other levels the codes focus on other features of the items (such as level of demand or order of skill). The literatures in this area informed the codes used; for example, some studies have suggested that the use of real-world contexts in questions can impact upon the achievement of different groups of students (Ahmed and Pollitt, 2001; Cooper and Dunne, 1998). This allowed the researchers to consider whether trends in DIF emerged across items with certain features.

Session B: Characteristics of Questions/Items

4. Question quality: The concept of quality in the context of exam questions

Victoria Crisp, Martin Johnson and Filio Constantinou (Cambridge Assessment, United Kingdom)

The importance of high quality questions in exams is often discussed in relation to validity and alongside the need to ensure accurate and reliable marking. Having good quality questions in exam papers is important to ensure that the intended knowledge, understanding and skills are assessed and to reinforce fairness. This is particularly pertinent at a time of reform of key qualifications in England, which means that new qualifications and sample assessment materials are being developed. Thus, it would seem important to have a good grasp of what makes good quality questions and, indeed, what is meant by 'quality' in the context of question writing. The latter is likely to be a socially-constructed concept and could relate to notions of validity at the level of the question. Previous studies have considered how the features of exam questions affect the ways students understand and react to a question and hence how difficult and fair the question turns out to be (e.g. Pollitt et al., 1985; Fan, Mueller, & Marini, 1994; Cooper & Dunne, 2000; Crisp & Sweiry, 2006; Ahmed & Pollitt, 2007). These have tended to involve quantitative and qualitative analyses of student performance on questions, and experiments involving manipulated versions of exam questions.

The current research adds to previous work by using views from question writers to explore the notion of 'quality' in exam questions and the features thought to contribute to the quality of questions. Seven exam question writers from four different subjects were shown some example exam questions. For each question the participants were asked to comment on the quality of the question and then to reflect on statistical data and senior examiner comments regarding how students performed on the question in the live exam. They were also asked to rate the quality of each question and to comment more broadly on their conception of what constitutes quality in exam questions.

Three conceptions of question quality emerged. Good quality questions:

- Test the intended knowledge, understanding and skills. Questions are clear around what is required, and lack ambiguity. This means that students understand the question and what they are being asked to do, allowing them to perform as well as they can. This theme relates to validity in terms of ensuring that the intended constructs are being tested.
- Differentiate between students who are better and weaker in the subject, whilst being accessible to all.
- Go beyond simple recall and understanding to assess skills in the subject and 'make students think'.

The discussions also identified a range of question features thought to affect quality. These included, amongst others: logical flow; clear and unambiguous; appropriate layout and spacing; clear resources; appropriate mark allocation; and a variety of task types or skills assessed. These are similar to some of the features identified as affecting difficulty and validity (at the level of the question) in previous research.

References

- Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: an experimental investigation of focus. Assessment in Education: Principles, Policy and Practice, 14(2), 201-232.
- Cooper, B., & Dunne, M. (2000). Assessing children's mathematical knowledge: Social class, sex and problem solving. Philadelphia, PA: Open University Press.
- Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. Educational Research, 48(2), 139-154.

- Fan, N., Mueller, J.H., & Marini, A.E. (1994). Solving difference problems: Wording primes co-ordination. Cognition and Instruction, 12(4), 355–369.
- Pollitt, A., Entwistle, N., Hutchinson, C., & De Luca, C. (1985). What makes exam questions difficult? Edinburgh: Scottish Academic Press.

5. How do question writers compose examination questions? Question writing as a socio-cognitive process

Martin Johnson, Filio Constantinou and Victoria Crisp (Cambridge Assessment, United Kingdom)

Setting written examinations has a long historical tradition, being highly prevalent across different education systems and learning contexts (e.g. schools, universities, and professional assessments) (Gilbert, 2011; Raban, 2008; Russell, 2002). Whilst the practice of writing examination questions is ubiquitous, and may even be considered to constitute a genre of writing, limited research has been carried out on how setters [question writers] write questions.

The limited previous research in this area has tended to adopt a cognitivist approach, using verbal protocol methods to explore what question writers think about when writing questions (e.g. Fulkerson, Mittelholtz and Nichols, 2009; Fulkerson, Nichols and Mittelholtz, 2010). This current project sought to extend this area of enquiry, using a variety of research methods, to look at question writing as a socio-cognitive process, so as to consider writing as both an internal and external process.

This project focused on the question writing practices of seven question setters from different subjects (Mathematics, Physics, Biology, Geography). The setters all had experience of writing questions for examinations based at the General Certificate of Secondary Education[1] level.
To capture evidence of the setters' writing practices each setter was observed remotely via video recording software whilst they carried out a question writing task. The visual data available to the researchers included a view of the setter's on-screen writing as well as their off-screen activity (e.g. note taking on paper, document use, etc.). The video material was then used to facilitate a stimulated recall session where individual setters could explain the processes that motivated their observed behaviours. All participants reported that the processes they had gone through under observation were representative of their usual practice.

Analyses suggested that the setters shared a common model of the writing process that paralleled a general cognitive model of writing proposed by Flower and Hayes (1981). This question writing model comprised three basic but interconnected phases: thinking about writing; writing and reflective thinking; and reviewing. In addition, question writing practice was influenced by the social system of examining through the way that the setters considered a variety of perspectives during the writing task. This insight gives an indication of how expert setter practice develops, implicating a sociocultural perspective which suggests that the broader social context of examination question writing is an inevitable influence on setter practice (Wenger, 1998).

References

- Flower, L., & Hayes, J.R. (1981). A cognitive process theory of writing. College composition and communication, 32 (4), 365-387.
- Fulkerson, D., Mittelholtz, D.J., & Nichols, P.D. (April 2009). The psychology of writing items: Improving figural response item writing. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

[1] The General Certificate of Secondary Education (GCSE) is an academic qualification awarded in a specified subject in England, Wales, and Northern Ireland. It is the most common qualification taken at the end of compulsory school education by students at roughly 16 years of age.

- Fulkerson, D., Nichols, P.D., & Mittelholtz, D.J. (May 2010). What item writers think when writing items: towards a theory of item writing expertise. Paper presented at the annual meeting of the American Educational Research Association, Denver, Colorado.
- Gilbert, I. (2011). Why do I need a teacher when I have Google? The Essential Guide to the Big Issues for Every 21st Century Teacher. Oxford: Routledge.
- Raban, S. (2008). Examining the World: A History of the University of Cambridge Local Examinations Syndicate. Cambridge: Cambridge University Press.
- Russell, D. R. (2002). Writing in the Academic Disciplines: A Curricular History. 2nd Ed. Carbondale, IL: Southern Illinois University Press.
- Wenger, E. (1998). Communities of Practice: Learning, Meaning, and Identity. Cambridge: Cambridge University Press.

6. Investigating experts' perceptions of examination question demand

Tom Bramley (Cambridge Assessment, United Kingdom)

The demand of exam questions is a concept that needs to be distinguished from empirical difficulty, intended difficulty and perceived difficulty. Demand is usually regarded as a qualitative concept. Pollitt, Ahmed & Crisp (2007) defined demands (plural) as '...separable, but not wholly discrete skills or skill sets that are presumed to determine the relative difficulty of examination tasks and are intentionally included in examinations.' One reason for investigating demand is in order to be able to predict empirical difficulty. If this prediction could be made accurately, then it could have application in standard-setting, standard-maintaining, pre-testing and item-generation contexts. Except in some very restricted domains, however, it does not yet seem that there is a good predictive relationship between a priori demand(s) and question difficulty.

However, in the UK context, the main reason for wanting to define demand has been for the purpose of making comparisons between individual questions and examinations in terms of their demand (or demands). For this purpose, demands have been conceptualised and operationalised through the CRAS scale. This is a tool by which expert judges make ratings of exam questions on a 5-point scale with respect to their Complexity, Resource requirements, Abstractness, Task Strategy and Response Strategy.

This study followed the approach to social science research known as Facet Theory (e.g. Borg & Shye, 1995) to formalise the definition of the concept of demand and to investigate the multivariate structure of ratings of exam questions using the CRAS scale. Ratings of short- and extended-answer exam questions in two contrasting subjects (English as a 2nd language, and Physics) were analysed with the relatively little known POSAC method (Partial Order Scalogram Analysis with base Coordinates). This can be thought of as a form of ordinal principal component analysis that creates a 2-dimensional representation of a higher-dimensional structure. Profiles of ratings across the 5 variables can be classified in terms of 'comparability'. Two profiles are comparable if one has the same value or higher (lower) than the other for every element. Two profiles are incomparable if one is higher on some elements and lower on others. The POSAC analysis positions profiles on a 2-dimensional grid such that the maximum number of comparable and incomparable relationships are preserved in 2 dimensions. The original variables (the CRAS variables) can then be interpreted in terms of how they partition the 2-D representation into distinct regions (e.g. Shye, 1985).

The clearest finding was that demand is an approximately unidimensional concept. All CRAS dimensions correlated strongly with overall demand, and the majority of pairs of questions were comparable in both A level Physics and IGCSE English as a 2nd language. In Physics there was also evidence of a consistent pattern in the second dimension, which reflected differences in the perceived Abstractness of the questions. These findings emerged despite relatively low inter-rater agreement on the raw ratings.

In conclusion, the POSAC technique seems a suitable and even useful method for investigating the internal structure of CRAS ratings of examination question demand. It takes account of the ordinal nature of the data, can help in understanding its dimensionality, and allows it to be visualised in 2-dimensional form.

References

- Borg, I., & Shye, S. (1995). Facet Theory: form and content. Thousand Oaks: CA: SAGE.
- Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of examinations syllabuses and question papers. In P. E. Newton, J. Baird, H. Goldstein, H.Patrick & P. Tymms (Eds.), Techniques for monitoring the comparability of examination standards.London: Qualifications and Curriculum Authority.
- Shye, S. (1985). Multiple Scaling. The Theory and Application of Partial Order Scalogram Analysis. Amsterdam: Elsevier.

Session C: Improving standards

7. Applying formal concept analysis in assessment: can it help mediate between socio-political and technical understandings of the meaning of exam grades?

Alex Scharaschkin (AQA, United Kingdom)

UK GCSE and A-level examinations are curriculum-embedded assessments with a strong emphasis on tasks that require extended, constructed responses, such as essays, drawings, spoken presentations, etc. Performance standards for these assessments are socially constructed and negotiated. Both subject content, and what is valued with respect to performance in a subject (summarised in so-called ‘assessment objectives’), are ultimately decided by government and government agencies, in consultation (to varying degrees) with various stakeholders, and hence also subject to the political cycle. Several organisations are licensed to develop and administer public examinations. They aim to produce assessment procedures that encode and operationalise these assessment objectives fairly.

But fairness entails that a ‘grade A in A-level physics’ (for example) should, in some sense, ‘mean the same thing’, across assessment procedures run by different organisations, and across examinations in close (e.g. adjacent) years.

Can the desire for a constant, definitive account of what ‘grade A in physics’ means (as would seem to be required to check on fairness) be reconciled with a qualitative, evolving, and necessarily fuzzy account of what a grade A performance actually ‘looks like’ that reflects professional understanding of the assessment objectives (and is arguably required by teachers and learners)?

This paper will argue that some of the models that have been applied to analyse meaning in fields such as formal semantics and knowledge representation can play a part in bridging this gap.

To cope with the complexities of socially-constructed notions of ‘subject domains’ and ‘standards’, such models should be able to deal with the natural fuzziness of concept boundaries. They should enable reasoning about counterfactual scenarios (what performances would have been like had the assessment procedure been different), to allow for discussions about reliability, validity and fairness. The models should also link to accounts of what assessors are doing cognitively when they categorise responses, for example theories such as the prototype theory of concepts (Rosch, 1973; Hampton, J.A., 2011; van Eijck and Zwarts, 2004).

As an example, the paper will examine the potential application of formal concept analysis (FCA) in UK-style public examination assessment. FCA (Ganter et al., 2005) is a mathematical approach to deriving a concept hierarchy, or formal ontology, from a collection (a so-called 'formal context') of objects and their properties.

The paper will suggest that modelling the outputs of assessment procedures as formal contexts, and valuation-categories (such as grades) as formal concepts within these contexts, may shed light on the extent to which assessments are operating as intended with respect to valuing students' performances. It will explore the prospects for using insights from basic and more generalised formal concept analysis in the design of high-stakes assessments, and for facilitating dialogue between subject experts, assessment experts, and policy makers.

References

- Ganter, B., Stumme, G., and R. Wille (2005) Formal concept analysis: foundations and applications. Berlin: Springer.
- Hampton, J.A. (2011). Concepts and natural language, in Belohlavek, R. and G.J. Klir (eds) Concepts and fuzzy logic. Cambridge, MA: MIT Press.
- Rosch, E. (1973) On the internal structure of perceptual and semantic categories, in T. Moore (ed), Cognitive development and the acquisition of language, 111-144. New York: Academic Press.
- Van Eijck, J. and Zwarts, J. (2004). Formal Concept Analysis and Prototypes. Workshop on the Potential of Cognitive Semantics for Ontologies (FOIS 2004), Turin, 3 November 2004.

8. Improving the maintenance of standards in England: evaluating a comparative judgement approach to awarding

Kate Kelly, Charlotte Stephenson, Neil Stringer and Faith Jones (AQA, United Kingdom)

To maintain standards over time, it is necessary to ensure that students of the same subject ability receive the same grades, regardless of which set of question papers they have answered. In English national examinations, vagaries in paper difficulty are compensated for through adjusting the grade boundaries, in a process known as awarding. However, current awarding procedures can only monitor changes in difficulty indirectly through students' performances, or through statistical measures which are only valid when successive cohorts differ only in the difficulty of the papers they sat. This assumption is unlikely to hold consistently in practice and is particularly challenged by large-scale social or policy change: for example, during education reform. A potential solution lies in utilising comparative judgement (CJ).

CJ allows measurement scales to be constructed through the iterative presentation of pairs of stimuli to a pool of judges, who must choose which of the pair best meets a given criterion. Applied to questions from a single examination, this methodology allows the difficulty of each question within the paper to be quantified. If questions from multiple papers are included, difficulty estimates can be obtained for each paper. This allows the difficulty of items from previous papers to be compared with items from a yet-to-be-taken paper. Changes in difficulty can thus be quantified, and related to changes in grade boundaries using item-facility test equating. This is a novel approach to equating, the details of which have been presented previously by Kelly, Stephenson and Stringer (2015).

Following on from a small-scale pilot (Kelly, Wheadon & Stringer, 2014), a larger-scale study is underway to evaluate this CJ-based approach to awarding and the use of item-facility test equating. The study consists of two conditions, with two groups of participants undertaking each condition: Chemistry PhD students and first year Chemistry undergraduates. The first condition aims to replicate the ideal scenario for awarding, in which items from two AQA A-level Chemistry past papers are ranked in terms of difficulty, without the potentially confounding effect of candidate ability. However, it is not known whether item difficulties can be ranked reliably without exemplification through candidate responses. In the second condition, although the task remains one of judging item difficulty, exemplification is provided for illustrative purposes. Although this condition risks confounding candidate ability and question difficulty, the advantage of increased reliability through including example responses may outweigh the disadvantage of potential bias. Including this condition allows the impact of exemplification to be assessed.

In both conditions, each judge is presented with 90 pairs of items and asked to decide which of the pair is the most difficult. All judgements are carried out via nomoremarking.com and include questions from both papers. When all judgements are completed, nomoremarking uses a Bradley-Terry model to estimate the item difficulty for each stimulus. Item-facility test equating can then be used to estimate the grade boundaries. The CJ-estimated grade boundaries can then be compared against the actual grade boundaries, and boundaries estimated using standard test equating.

This presentation will report on the findings of the above study. This work has potentially valuable applications: if it is technically credible, CJ-based awarding would present a more valid and robust alternative to current procedures. Directly quantifying difficulty is more philosophically coherent, and arguably fairer: unlike judgements of script quality, CJ difficulty estimation should not be susceptible to any student-induced biases such as the Good and Cresswell effect (Good & Cresswell, 1988) and the consistency bias (Scharaschkin & Baird, 2000). In avoiding the assumptions inherent in current statistical methods, a CJ approach could also help overcome the standards maintenance challenges presented by changing educational policies.

9. Assessing Primary Writing in a Politically High Stakes Context

Sarah Maughan (AlphaPlus, United Kingdom)

Assessment design and development is generally based on decisions about the most valid means of assessing the target construct in a way that will be both reliable and manageable. The aim is to develop the best possible test. However, this approach ignores the social and political decisions that also need to be taken into account in the design of an assessment. Choices about the right balance to strike between validity concerns and reliability concerns are hugely impacted by the stakes of the assessment: assessments designed for a low stakes context, such as to inform teaching approaches, could be designed to be highly valid – assessing contribution to discussion for example, whereas it is often decided that such an approach would be too risky for a high stakes context.

The national primary school tests in England are a very high stakes endeavour. The results from end of primary school tests, for children aged 11-years-old, are provided to parents, children and secondary schools as an indication of primary school attainment. More importantly, however, the combined results for the entire year group are used to judge how well the school has served that cohort of children. Action may well be taken against schools that have not reached target levels of attainment.

Until 2012, children were tested on their mathematics, reading and writing skills at the end of primary schooling. However, the Bew Review (Bew, 2011) recommended that the externally-set writing assessments should be removed and replaced by a teacher assessment of extended writing attainment and an externally-set test of grammar, punctuation, vocabulary and spelling (GPS). The reason for the recommended change was the subjective nature of the writing external marking and the unreliability inherent in it. In short, a more reliable assessment was required in the context of a very high stakes accountability system. This change was introduced in the summer of 2013.

The new GPS tests could be considered to be a compromise – so the more valid extended written test has been replaced with a more reliable test form. Given the context, it is essential that the test items are as close to a proxy measure for writing skills as possible. This presentation will demonstrate some GPS items developed for an educational publisher in England (Rising Stars) with an analysis of how well they reflect the wider construct of 'writing'. A range of item types will be included which have been designed to assess different 'writing' skills.

References:

- Lord Bew, Independent review of key stage 2 testing, assessment and accountability: final report, 23 June 2011 [Available: https://www.gov.uk/government/publications/independent-review-of-key-stage-2-testing-assessment-and-accountability-final-report, April 2016].

Session D: Supporting teacher's and rater's assessment practices

10. Exploring Teachers' Approaches to Classroom Assessment: An Instrument Development Study

Christopher DeLuca, Danielle Lapointe-McEwan, Adelina Valiquette and Andrew Coombs (Queen's University, Canada)

Purpose

Teacher assessment literacy has emerged as dominant requirement within the current accountability framework of educational systems throughout the world (Brookhart, 2011; Stobart, 2008). Recent policies throughout Europe, North America, Australia, and New Zealand have emphasized classroom teachers' use of formative and summative assessments to guide instruction and support student learning (Birenbaum et al., 2015). However, despite this priority, research shows that teachers struggle to interpret contemporary assessment policies, with significant variability in teachers' implemented assessment practices (Bennett, 2011; MacLellan, 2004). Furthermore, previous assessment literacy instruments used to measure and support teachers' assessment practices, maintain weak reliability and validity evidence in relation to contemporary professional assessment standards (Gotch & French, 2014). The purpose of this paper is to construct a reliable instrument to measure teachers' approaches to assessment in relation to current classroom assessment standards (i.e., JCSEE, 2015).

Methods

Following Plake et al.'s (1993) protocol for generating the Teacher Competencies Assessment Questionnaire, we developed the Approaches to Classroom Assessment Inventory (ACAI) based on the recently published Classroom Assessment Standards (JCSEE, 2015). Using an expert-panel method to collect validity evidence, the questions on the ACAI were analyzed for their alignment to the Standards by 10 North American assessment experts and 10 teachers (5 elementary and 5 secondary). Based on feedback from experts, revisions were made to the instrument. The final inventory included four sections: (a) five scenario-based questions that determined teachers' approaches to assessment across four assessment literacy themes: assessment purpose, assessment process, assessment fairness, and measurement principles; (b) 26 Likert-scale questions to measure teachers' confidence in various assessment tasks; (c) 26 Likert-scale questions about teacher' professional learning priorities and preferences, and (d) demographic questions. The instrument was administered to 400 teachers from across the Canada to collect reliability and validity evidence. Analysis was conducted using descriptive statistics, ANOVAs, and factor analyses. Reliability statistics were then calculated for resulting factors.

Results
In the complete paper results, we present our entire instrument development process as well as our empirical results. For this proposal, we present initial descriptive results based on teachers' responses to the first section of the ACAI.

The first section of the ACAI provided teachers with 5 assessment scenarios asking them to select their priority response based on a list of options. For each scenario, the response options reflected three dimensions within each of the assessment literacy themes: (a) assessment purposes (dimensions: summative, formative/assessment for learning, and assessment as learning); (b) assessment processes (dimensions: designing assessment, scoring and administering assessments, and communicating assessment results); (c) assessment fairness (dimensions: standard approach, equitable approach, and differentiated approach); and (d) measurement principles (dimensions: reliability, validity, reliability and validity). Results indicate that teachers overwhelmingly prioritize a formative assessment approach (84%) over a summative approach (4%). For assessment processes, teachers identified that they adjusted 'scoring and administering practices' to deal with scenarios compared to 're-designing assessments' and 'communicating assessment results differently.' 60% of teachers use an 'equitable approach' to grading/assessment (i.e., based on exceptionalities) with 37% of teachers differentiating assessments for diverse learners. Finally, the majority (71%) of teachers prioritized validity considerations over reliability considerations.

Significance
Measuring and supporting teachers' assessment literacy has been a focus of educational policy and research since the early 1990s. Through this study, we have constructed a contemporary instrument with initial reliability and validity evidence to measure teachers' approaches to assessment. We see value in continuing to build additional validity evidence for sound measures that accurately characterize teachers' strengths and weaknesses in assessment. These measures can then form the basis for responsive teacher education that works to enhance teachers' classroom assessment practices.

11. Supporting Teachers' Assessment Professionalism in the Context of Change in Curriculum and Examinations

Kay Livingston, Louise Hayward, George MacBride, Carolyn Hutchinson and Ernest Spencer (University of Glasgow, United Kingdom)

This paper discusses some of the complexity of and potential action to develop assessment professionalism in the context of research on quality in teachers' continuing professional development generally (e.g. Caena's, 2011 Literature Review for European Commission) and on effective promotion of assessment for learning through learning communities (e.g Hill's, 2016). Different kinds of policy thrust and different kinds of guidance about assessment practice create tensions for teachers within many countries. While teachers may be professionally drawn towards ideas about assessment fully integrated in learning, their teaching and assessment practice may be strongly influenced by external examinations or tests, which can significantly affect students' life chances and teachers' and schools' professional standing. Such potentially conflicting pressures are particularly prevalent when there is change in curricular and assessment arrangements and where standard setting and improving attainment are policy features. They constitute the context in which both implementation of new policy and teachers' continuing professional learning happen.

The paper uses case study evidence from interviews with 80 16-year old students and 80 teachers representing almost all subjects in 5 secondary schools in different parts of Scotland. One key aim was to analyse early experience and understanding of transitions from the Broad General Education Phase (ages 3-15) of the Scottish Curriculum for Excellence, the policy rhetoric for which strongly emphasises assessment for learning, to the Senior Phase of newly implemented National Qualifications. Separate national agencies provide assessment guidance for Broad General Education and senior examination specifications.

Despite frustration with the range and quality of support for assessment provided nationally and locally, the teachers wanted to develop their assessment professionalism – to access good exemplification of national standards, participate in discussion with colleagues in their own school and elsewhere about them and achieve both effective assessment for learning for their students and accuracy and consistency of judgements of overall achievements. Washback effects of Senior Phase qualifications were equally evident. Teachers tended to use traditional testing approaches for summative assessment, often modelled on examinations and justified on the grounds that these approaches are more rigorous and reliable.

The study identified action, including further research, required specifically in the Scottish context, but also raised general issues relevant to countries undergoing processes of curriculum and assessment change.

The paper discusses two questions of general relevance:

- What are the potential implications for teachers' assessment professionalism when national policy and practical advice derive from two (or more) sources/organisations?
- How can a 'best fit' approach to assessment of achievement through teacher professional judgement of a portfolio of classwork be effectively developed and moderated when teachers believe that testing and grading are more rigorous – are there potential advantages in incorporating teacher-designed tests in 'best fit' approaches?

Implications and potential action for professional learning are discussed, taking account of teachers' existing beliefs and practice and drawing particularly on situated learning theory (Wenger, 1998, 2010) and the effective approaches to professional development identified by Caena (2011) and Hill (2016). The paper also proposes ways of achieving better synergy between assessment for learning advice and national guidance about sharing standards and moderation.

References

- Hill, M. F., 2016. Assessment for Learning Community: Learners, Teachers and Policymakers. In Wyse, D., Hayward, L. and Pandya, J. (Eds) The SAGE Handbook of Curriculum, Pedagogy and Assessment. London: SAGE Publications.
- Caena, F, 2011. Literature Review: Quality in Teachers' Continuing Professional Development. European Commission Report.
- Wenger, E. 1998. Communities of practice: meaning and identity. Cambridge: Cambridge University Press.
- Wenger, E., 2010. Communities of Practice and Social Learning Systems. In Blackmore, C (Ed) Social Learning Systems and Communities of Practice. London: The Open University with Springer-Verlag.

12. Using comparative judgement for the assessment of academic writing: an examination of its validity

Tine van Daal, Marije Lesterhuis, Liesje Coertjens, Vincent Donche and Sven De Maeyer (University of Antwerp, Belgium)

Theoretical framework

Nowadays, essays are mostly assessed using rubrics. As essays are open-ended, it is harder to score them in a reliable way. Furthermore, since it is impossible to do justice to the complexity of a competence as academic writing with a list of pre-defined criteria, absolute analytic judgement using rubrics has problems regarding its validity (e.g., Sadler, 2009). Using comparative judgement (CJ) judges do not have to rely on predefined criteria, but compare two pieces of work and then decide holistically which piece of work is better regarding academic writing. Up to now, this method shows to be very promising in reaching reliable scores. Furthermore, CJ is claimed to be a valid approach for the assessment of writing (Pollitt, 2012ab). Thereby, its validity is rooted in its holistic character and in the extent to which the final rank-order reflects the shared consensus on what a good academic essay comprises. These assumptions are, however, not backed by empirical evidence. To test their holding, critical implications arising from both assumptions should be identified and investigated (Kane, 2013).

Regarding the first assumption, implications refer to the extent to which CJ allows and even values variation between judges in their conceptualisation of good writing. This implies that CJ expects judges to differ in the focus and in the broadness of their judgements. Furthermore, the holistic character of CJ is also claimed to enable judges to tap into their expertise regarding good academic writing (Jones et al., 2015). With respect to the second assumption, rooting CJ in the shared consensus across judges assumes that the construct of academic writing is fully covered by the whole of judges' conceptualisations. This implies that all dimensions of academic writing should be taken into account (Messick, 1989). None of these implications are, however, backed by empirical evidence.

Methodology

With this study, we examine whether two assumptions regarding CJ hold within the context of the assessment of academic writing in a pre-master course at a Belgian university.
The implications arising from these assumptions are tested. Judges were only given a general competence description and did not receive any training.
After each comparison, judges were asked why they choose one essay above the other. These justifications were coded qualitatively and quantified to provide insight into the dimensions of academic writing judges took into account.

Results

Results show that almost 70% of the arguments are directly related to the competence description. As all dimensions of academic writing are covered, full construct representation is warranted. However, the weigh given to the different dimensions of academic writing varies. Furthermore, some of the arguments given are related to academic writing but not associated with the competence description. This type of arguments evidences that judges also use their expertise regarding academic writing while judging the essays. Additionally, it was found that judges differ in how they conceptualize academic writing. Judges' arguments vary in focus (type of dimensions looked at) and in broadness (number of dimensions simultaneously looked at). Furthermore, variations between judges were found in the extent to which they tap into their expertise. Finally, this study found some indications that misfitting judges (judges not conforming to the shared consensus) differ in conceptualisation.

Implications

The results of this study point at the significance of differences between judges' conceptualisation and the extent to which their expertise is tapped. Future research should replicate these findings within other contexts and gain more insight in which characteristics of judges cause these differences, how these differences influence the validity of the resulting rank-order and their relationship with misfit of judges.

Session E: Conceptions of subject difficulty and subject/test taking strategies

13. Inter-subject comparability: How does adjusting grade boundaries affect schools, subjects and candidates in England?

Caroline Lau, Simon Eason, Ben Jones (AQA, United Kingdom) and Mike Cresswell (Independent, United Kingdom)

Inter-subject comparability is the most problematic of the standards comparability dimensions, from both a philosophical and practical point of view. There is an extensive literature on both aspects (Newton et. al., 2010, Cresswell M J, 2010) but less on the practical outworkings, effects and implications (Lamprianou, 2009), were an agreed method for 'correcting' the perceived misalignment of subjects implemented. One of the main such practical outworkings is the use of aggregated subject grades for the purposes of school accountability measures. Currently, in England inter-subject comparability has been the subject of recent discussion (Ofqual, 2015a) as centres whose students take purported 'easier' subjects will, it is argued, score more highly on the high stakes measures (in England the imminent 'Progress 8' measure).

This paper will add to the inter-subject literature by looking at the practical effects at the student, subject and school level in England if a method for correcting perceived misalignment of subjects were implemented. Using comprehensive, national GCSE result data from England in 2015, subject grade boundaries were adjusted according to their relative grade difficulty as calculated using a Rasch analysis (Ofqual, 2015b) and students' marks were re-graded. Student level attainment values were then calculated based on original and adjusted grades. These values were then used to calculate an average attainment for each school. For the subject level analysis, subject pairs tables were produced for a range of adjusted subjects using both the original and adjusted grades.

A number of attainment measures were calculated to see how bringing the subject grades 'in line' would affect the rankings of the individual candidates and schools. In both cases, the changes in grade boundaries had little or no effect on the rankings, indicating that – even if philosophically defensible – the complex task of aligning subjects statistically may not be a worthwhile effort where these two parties are concerned. When looking at the changes at subject level, there was often little movement towards the 'centre line'; indeed, in some cases a very large shift away from it. This evidence suggests that it may not be possible to make subject grades interchangeable, and actually detrimental in some cases to try to do so.

This paper will be of interest to researchers, policy-makers and those involved in inter-subject comparability.

References

- Newton, P.E. (2010) Thinking about linking. Measurement: Interdisciplinary Research and Perspectives 8(1), 38–56
- Cresswell M J (2010) Defending the Quality of Links Between Scores from Different Tests and Exams, Measurement: Interdisciplinary Research and Perspectives 8(4), 157–160
- Lamprianou, I. (2009) Comparability of examination standards between subjects: an international perspective, Oxford Review of Education, 35(2), 205–226.
- Ofqual (2015a) Comparability of Different GCSE and A Level Subjects in England: An Introduction (ISC Working Paper 1, Coventry, Ofqual)

- Ofqual (2015b) Inter-Subject Comparability of Exam Standards in GCSE and A Level (ISC Working Paper 2, Coventry, Ofqual)

14. Subject Entry Choices and Perceptions of Subject Difficulty: Are the Two Linked, and if so, How?

Benjamin Cuff (Ofqual, United Kingdom)

Concerns have been raised about a disparity in the difficulty of different subjects studied in the UK, and that this might be contributing to a lower uptake in certain 'key' subject areas. However, debate continues over how best to conceptualise subject difficulty, and whether there truly are any measurable differences between subjects (as discussed in a series of working papers recently published by Ofqual, 2015). The purpose of the current research was to explore whether teachers' and students' perceptions of subject difficulty might be having an effect on which subjects students choose to study in secondary education, and whether other concerns (e.g., subject enjoyment or usefulness) might interact with, or supersede, perceptions of difficulty.

A qualitative research design was chosen to allow for an in depth exploration of these issues. Twelve schools from a variety of school types and geographical locations were recruited. Interviews and focus groups were held with 49 teachers and 112 students respectively. Thematic analysis was performed on coded transcripts and the main drivers of students' and teachers' behaviours were identified.

Results suggested that teachers were split in their perceptions. Some believed that some subjects are more difficult than others, whilst others believed in 'subject equality', because each is difficult in their own right. Nevertheless, all teachers agreed that whether a student found a certain subject difficult or not was very much dependent upon that student's individual strengths. Teachers had an influence over students' subject choices in two main ways: through the setting of policies (e.g., entry criteria and option blocks) and by giving advice. Entry criteria (e.g., prior attainment requirements) were often based upon notions of subject difficulty, which served to prevent students from taking subjects they would find too difficult. Some schools also chose not to offer certain subjects because they were seen to be too difficult, again preventing uptake in those areas. When advising students on their subject choices, teachers sometimes discouraged students from taking subjects that might be too difficult for them, but this was mostly done according to each student's individual strengths, rather than any societally held notions of subject difficulty. Although subject difficulty was an important consideration for teachers, a large part of their advice was based upon what each student would enjoy and find useful for future education or employment.

Students generally agreed that although some subjects stood out as being more difficult than others, whether or not they found a subject difficult or not was dependent upon their individual strengths. Students did base their subject choices on their personal perceptions of subject difficulty, and recognised that they were also occasionally discouraged by their teachers, parents, and friends from choosing subjects that were thought to be difficult. However, in line with teachers' views, students stated that perceptions of difficulty were not the main basis of their decisions, and focussed more upon enjoyment and usefulness. Importantly, students often stated that they were willing to overlook the difficulty of a subject when they enjoyed it and/or needed it to satisfy their future ambitions.

The main conclusions drawn from these results were that although perceptions of subject difficulty and subject entry choices do seem to be linked, perceptions of enjoyment and usefulness interact with, and often supersede, this relationship. In order to address those concerns raised about national skills deficits in certain subjects, and to encourage greater student uptake in those areas, it may be possible to enhance students' enjoyment and ambition in key subject areas.

References

- Ofqual, (2015). Inter-subject comparability: Research documents. Retrieved from https://www.gov.uk/government/collections/inter-subject-comparability-research-documents

15. Test Taking Practices, Background Variables, and their Relationship to Validity

Elena Papanastasiou and Agni Stylianou-Georgiou (University of Nicosia, Cyprus)

According to the Standards for Educational and Psychological Testing (2014), validity is the most fundamental consideration in evaluating tests. However, construct-irrelevant variance is a major threat for validity. Construct-irrelevant variance may arise from numerous sources, such as lack of clarity in instructions, lack of consistency in test preparation, or by item or test formats (including the use of bubble sheets) that might be unfamiliar to certain populations. However, although research has examined these issues in regard to high stakes test, this is not necessarily the case for low-stakes tests such as in international studies.

In countries such as the USA with large testing cultures, students take many multiple-choice high-stakes tests starting from elementary school. Due to the high-stakes nature of these tests, teachers typically expose their students to test taking strategies (TTS), such as whether answer changing should be used, or whether answers should be omitted when unknown. However, such instruction cannot necessarily generalize to students in other countries with low testing cultures, or with tests that mostly include open-ended questions. Therefore, this study serves as a starting point to determine whether background factors differentiate elementary school student's TTS, in a first attempt to determine whether such differences could be a threat to a study's validity. The main purpose of the current study was a) to determine variables that could explain part of the variation in student's test-taking strategies, and b) to determine whether some student's exposure to the TIMSS study has differentiated their use of TTS.

The sample of this study is composed of 291 fourth-graders in Cyprus (which has a low testing culture), of which 45.3% were male. The students were administered a 30-item questionnaire that measured frequency of using test-taking strategies, attitudes, and background questions, The data were analyzed with the use of descriptive and inferential statistics (e.g. MANOVAs). The results have shown that differences do exist in certain TTS among groups of students. For example, students who had participated in academic competitions, and thus had more testing experience, were least likely to use answer-changing. However, students who had participated in TIMSS reported using answer-changing more frequently in occasions where they had reread and better understood the questions. This study also found that the students whose native language was not Greek, or who had not participated in academic competitions had more negative attitudes towards multiple-choice tests.

In conclusion, this study found that despite the homogeneous student culture in Cypriot elementary schools, statistically significant differences were found in student's attitudes towards multiple-choice tests, and in regard to the TTS of answer-changing. One would expect, that the differences might be larger when compared to students from other countries, as was found in a PISA study where the highest achieving countries were the ones with the least amount of omitted response rates (Gillmore, Longback & Poggio,2014).

The significance of this study lies in the finding that background factors are related to TTS which could increase the test error variance. Therefore, from a research perspective, these results strengthen the need for the replication of this study for between country comparisons, with the administration of achievement tests to measure the actual error variance and their effects on validity. From an educational perspective, these results also identify the importance of uniformly familiarizing students with the testing process and with TTS in order to minimize the effects of backgrounds variables that are not construct related.

References

- Author, (2014). Standards for educational and psychological testing. Washington DC: American Educational Research Association.
- Gillmore, S., Longabach, T., & Poggio, J. (2014). A new threat to validity: An examination of cultural discrepancies in omission rates on international assessments. Paper presented on the AERA conference, Philadelphia,PA.

Session F: Impact of educational policy on equity and social justice

16. An exception that proves (tests) the rule: social justice and national standardised assessment policy in Scotland

Louise Hayward, Kay Livingston, George MacBride and Ernest Spencer (University of Glasgow, United Kingdom)

Many European education systems, regardless of the political parties in power, are characterised by national standardised assessment systems and data gathering processes (Eurydice, 2012); concerns about the unintended consequences of such processes have been widely expressed (Eurydice, 2009, Mons, 2009). Many European governments also have policies to address issues of social justice in education (Schraad-Tischler, 2015). Alignment between these two policy strands is often tenuous or absent.

Scotland affords a case study to explore issues related to the alignment of national assessment policy with the promotion of social justice. Scotland is an outlier in the European context: national testing has been and remains contentious and national data gathering processes have been relatively light touch. Scotland is also an outlier as a country marked by high levels of inequality with notable consequences for equity in school attainment (Authors, forthcoming, OECD, 2007); governments have consequently afforded social justice and educational inclusion a very high priority. Recently, the First Minister reversed policy, introducing national standardised testing on the grounds of promoting social justice (Scottish Government, 2016). The ‘outlier’ nature of the case study facilitates identification and analysis of issues common to education systems across Europe but which are here particularly overt.

The authors draw on international, national and local research and critically analyse evidence provided by government policy documentation, political party manifestos, teacher professional association policies and publicly expressed views of practitioners.

The authors firstly outline the political, social and educational contexts in Scotland which informed widespread opposition to national standardised testing in primary and early secondary education over two decades, leading ultimately to its removal. The paper proceeds, through the lens of social justice, to describe, analyse and critique the proposals to reintroduce national standardised testing. The authors identify reasons for this reversal of long-standing policy and, in the context of existing European research, provide a critique of the validity of the proposed assessment system and of its likely impact on social justice.

The authors explore the potential results and impact on the stated aim of any government to promote social justice when policy makers, not only in Scotland but in many European countries, fail to address effectively tensions inherent in national assessment systems, between for example:

- (contested) claims for the value of teachers’ professional judgement and (apparently widely accepted) claims for the reliability of external assessments
- informed recognition of the complexities of valid assessments and demands for readily available statistical data
- claims of the value of curricular breadth and depth of learning and claims of the importance

of assessing core knowledge
- claims that national assessment aims to inform support for learners and the development of accountability processes
- the intended aims of policy and implementation in practice
- the intended aims of policy and impact on learners.

They consider implications for policy makers, teacher professional associations and the research community if such tensions are to be addressed and conclude by proposing key criteria which a national assessment system must meet if it is to promote social justice in education.

References

- Authors (forthcoming) Scotland: The intersection of international student assessment and educational policy development in Volante, L (ed) (forthcoming) The Intersection of International Achievement Testing and Educational Policy. Abingdon: Routledge
- Eurydice (2009) National Testing of Pupils in Europe. Brussels: EACEA
- Eurydice (2012) Key Data on Education in Europe 2012. Brussels: EACEA
- Mons, N (2009) Theoretical and real effects of standardised assessment. Brussels: EACEA
- OECD (2007) Quality and Equity of Schooling in Scotland. Paris: OECD
- Schraad-Tischler, D (2015) Social Justice in the EU – Index Report 2015. Bertelsmann Stiftung
- Scottish Government (2016) National improvement framework for Scottish education. Edinburgh

17. School characteristics, SES and achievement

Trude Nilsen (University of Oslo, Norway) and Jan-Eric Gustafsson (University of Gothenburg, Sweden)

A major goal in educational policy is to promote equity among students. Nevertheless, the strong relation between students' socio-economic status (SES) and achievement has been established in numerous studies and continues to persist (OECD, 2013). Hence, school factors that reduce this relation need to be identified. However, little is known about which school factors influence the relationship between SES and achievement.

School factors that are important for student learning and achievement have in some studies been found to mediate and moderate the relation between SES and achievement. School climate creates the foundation for instruction and learning and has been found to influence the relation between SES and achievement (e.g. Uline & Tschannen-Moran, 2008). Instructional quality (cognitive activation, supportive climate, clarity of instruction, and classroom management) and instructional quantity (time allocated to instruction) have been shown to positively influence student achievement and to both mediate and moderate the relation between SES and achievement (e.g. Rjosk et al., 2014; Willms, 2010). There are, however, few studies investigating the relation between school factors and equity, especially using moderation models.

We hence pose the following research question: To what extent can the within-school relation between SES and mathematics achievement be moderated by school characteristics reflecting quality and quantity of instruction, and school climate?

Methods

Data from the 50 countries (N = 282 737 students) that participated in grade 8, TIMSS 2011, was analyzed using Mplus 7.3. Two-level (students and schools) random slopes, multiple-group structural equation models were specified to investigate whether school characteristics moderate the within-school relationship between student SES and achievement.

Students' ratings were used to measure SES (a scale derived from students' ratings of the number of books at home, their parents' highest education and home study supports) and Instructional

Quality (e.g. 'I know what my teacher expects me to do', and 'My teacher is easy to understand'). Principals' ratings were used to measure instructional quantity (time allocated for instruction) and school climate (School Emphasis on Academic Success (SEAS) and a safe and orderly climate). The Human Development Index (HDI) was added to the dataset at the country level.

Results and discussion

The school factors reduced the strength of SES in 22 cases, including mainly high HDI countries. However, the strongest determinant was School-SES. The results varied across countries, and for some, more than one school factor reduced the strength of SES. In general, the relation between students' SES and their achievement was weaker in highly developed countries if: school-SEs was high; schools had higher instructional quality; allocated more time to instruction; put more emphasis on academic success; and/or had a safe and orderly school climate. The school characteristics thus predicted higher equity in these countries. In some countries though, the school characteristics enhanced the strength of SES. This inequity could be due to unequal distribution of access to high quality education across different social groups of students. For instance, schools with high SES may have high quality teachers with excellent instruction.

Our findings may contribute to educational policy in that high quality and quantity of teaching as well as school climates that emphasize academic success and are safe and orderly, may promote equity among students.

References

- OECD (2013). PISA 2012 Results: Excellence through equity. PISA, OECD Publishing.
- Uline, C., & Tschannen-Moran, M. (2008). The walls speak: The interplay of quality facilities, school climate, and student achievement. Journal of Educational Administration, 46(1), 55-73.
- Willms, J. D. (2010). School composition and contextual effects on student outcomes. The Teachers College Record, 112(4), 3-4.

18. Teacher quality mediating and moderating the relation between SES and achievement in Nordic countries

Hege Kaarstein, Trude Nilsen (University of Oslo, Norway) and Jan-Eric Gustafsson (University of Gothenburg, Sweden)

In a global perspective, the Nordic countries are known to have small differences between schools and educational policies that promote equity. However, these societies are changing. Internationalization including international large-scale assessment may incite a washout of the characteristics of the Nordic countries. Differences between schools seem to increase (Kjærnsli & Olsen, 2013) and particularly so in Sweden (Yang-Hansen, Gustafsson & Rosén, 2014).
The relation between socio-economic status (SES) and achievement is one common indicator of equity, and there are indications that the strength of this relation is increasing. Yet, we know little about the mechanism through which SES is related to educational achievement. If the aim is to reduce the effect of SES on achievement, school characteristics that reduce this relation and that are within the control of educational stakeholders need to be identified.

There is a general agreement that teachers matter more than any other school factor and that teacher quality (TQ) is the foundation for effective schools, learning and instruction. Some studies also indicate that TQ may play a role for equity (e.g. Darling-Hammond, 2006). TQ may exert both additive effect (i.e. mediating the relation between SES and achievement) and interactive effects (i.e. moderating the relation between SES and achievement).

We address the following research question:
Does TQ mediate and/or moderate the relation between SES and achievement in Nordic countries?

Method
Data from Norway, Sweden and Finland (N = 13 345 students) that participated in grade 8 in Trends in International Mathematics and Science Study (TIMSS) 2011 was analyzed using Mplus 7.3. Two-level (students and schools) mediation and moderation structural equation models were specified using a multiple-group approach.

TQ was measured by teachers' ratings of items pertaining to: experience, educational level, major area of study, job satisfaction, confidence, self-efficacy and professional development. An index for SES based on students' ratings of a number of items in the home was used.

Results and concluding remarks
Several interesting preliminary findings have been found. Moderation models indicate that Norwegian teachers whose major area of study is mathematics education, and teachers who report being very well prepared to teach algebra, contribute to reduce the influence of SES on achievement. Hence, in Norway, teachers who are well skilled in mathematics and mathematics education seem to promote equity among students. In contrast, for Finland the results indicate that teachers who are well skilled in mathematics exacerbate inequality. The results from mediation models may shed light on this finding, given that they indicate a positive association between school-SES and teacher education. This suggests an unequal distribution of access to high quality teachers across different social groups. In Sweden, results from the moderation models show that only teachers' job satisfaction is related to inequity, and the mediation models indicate that teachers more often participate in professional development in high SES schools.

In summary, teachers with high mathematics competence as well as the distribution of these teachers to low versus high SES schools seem to play a role for equity. In addition to contributing to the field of equity and applied methodology, these findings have implications for educational policy and will be discussed in light of previous research and current policy trends.

References

- Darling-Hammond, L. (2006). Securing the Right to Learn: Policy and Practice for Powerful Teaching and Learning. Educational Researcher, 35(7), 13-24.
- Kjærnsli, M. & Olsen, R. V. (2013). Norwegian PISA 2012 report. Oslo: Universitetsforlaget
- Yang Hansen, K., Gustafsson, J.-E., & Rosén, M. (2014) School Performance Differences and Policy Variations in Finland, Norway and Sweden. In Northern Lights on TIMSS and PIRLS 2011. Denmark: Nordic Council of Ministers.

Session G: Measurement of complex skills

19. The Primary Scientific Reasoning Test – In Pursuit of Content Validity

Diana Ng (Oxford University Centre for Educational Assessment, United Kingdom)

Attention to the development and importance of scientific and reasoning skills to learners of science has significantly intensified over the last two decades. Competence in science, including well-developed reasoning faculties, is valued in a globalised era dominated by complex scientific issues and sophisticated technological artefacts. Science education programmes in many countries, including Singapore and the UK, have responded by actively promoting teaching approaches and frequent curricular-policy reviews that foster the desired reasoning and cognitive skills in pupils.

Despite a large body of empirical findings on particular aspects of reasoning skills utilised in science tasks, it is still unclear how scientific reasoning could best be assessed or if existing assessment tasks are adequately measuring this higher-order cognitive proficiency. Progress in defining criteria for measuring scientific reasoning is largely constrained by a limited understanding of the specific nature of scientific reasoning, which in turn impacts the development of suitable instruments. This state persists despite the existence of multiple proposed frameworks for scientific reasoning and a long-standing interest in assessments, much of which has not been systematically tested. Notably, there is limited information produced about the relationship between scientific knowledge and scientific reasoning. A better understanding of the relationship could help educators evaluate the impact of the prescribed curriculum on the development of reasoning abilities, particularly as many current scientific reasoning tests do not require test takers to use scientific knowledge when responding to the tasks or items. Instead the relevant content is provided in the tasks or items. Minimising the influence of content while assessing pupils' reasoning is known as a knowledge-lean approach, which has been criticised on the grounds that content knowledge can impact scientific reasoning or is part of scientific reasoning.

The present study investigates the content validity of a paper and pencil test – the Primary Scientific Reasoning Test (PSRT), when tested on school children in Singapore. This study is part of research developed to conceptualise and design items for the PSRT, as well as examine the validity of the scientific reasoning construct of the test with children in England and Singapore who have just completed their primary education. The scientific reasoning construct of the test is based on commonalities between emerging philosophical and psychological perspectives. Unlike knowledge-lean tests, items in the PSRT require test takers to draw on learnt content knowledge.

This presentation reports on the methodology employed and preliminary outcomes of the content validity of the PSRT using Sireci's framework (1998). In this framework, content validity is defined as the 'degree to which a test measures the content domain it purports to measure' (p. 299) and is described as being composed of four features of test quality. These features are domain definition, domain representation, domain relevance, and appropriateness of the test development process. This presentation will show items that exemplify the sequential process to gather evidence of content validity, beginning with the item design level and followed by feedback from a Singapore-UK-US expert panel, a study with Singapore pupils, and the iterative revision of the items. The presentation will contrast the items constructed in this manner with the items in a current knowledge-lean test of formal reasoning – the Classroom Test of Scientific Reasoning (Lawson, 1978, 1995). The comparison will highlight specific features of the tests and items that purport to assess higher order reasoning skills, thereby facilitating a better understanding of the relationship between science content and scientific reasoning. Finally, the presentation will discuss the significance of pursuing content validity in reasoning tests at various levels – conceptual, epistemic, methodological and pedagogical, with potential impact on policy.

20. Multistage testing and Disability Act: a new test method for policy evaluation

Reinaldo Dos Santos and Thierry Rocher (French Ministry of Education, France)

In France, the Act on Disability of the 11th of February 2005 makes it mandatory to ensure the enrollment of every child with disabilities in the standard school system, as close as possible to one's home, to ensure continuity of one's curriculum and ensure equal opportunities for exams. Ten years later, time has come for an evaluation of this policy: do disabled children benefit from this intended inclusive education policy?

Within this context, a large-scale longitudinal survey was set up in 2015: a sample of 5895 children with disabilities born in 2005 was randomly selected. The age was chosen in order to respect two conditions: 1-those children never went to school before the Act, as they are 11 years old; 2-the majority of these children are at the end of the primary cycle in French schooling.

The design of this survey includes a standardized assessment of these children in terms of cognitive skills, aiming to measure how they progress in their curriculum. But this project has raised many questions, especially this: what is the best way to assess a population that is really heterogeneous, because of the wide diversity of disabilities? How can we transform our traditional assessment tools to suit the disability of each child, with the least possible bias?

These questions led us to quickly consider new methods of assessment.
The diversity of the target population being such as it prohibits the use of a conventional linear test, we took the turn of implementing adaptive testing. Because the full adaptive test, with a reorientation of the test-taker after each answer, seemed inappropriate to us, for practical reasons, but also because of unnecessary complexity, we chose a Multi Stage Testing model (Yan, Davier and Lewis, 2014), with six different levels and three degrees of depth.

We chose to implement a computer-based assessment in reading and mathematics, with mainly closed questions, which could be scored automatically so that the test taker can go further in the test tree. We chose items from our national item bank. These items are calibrated on large-scale samples, at different grades in primary schools. Thus, we could use items parameters to draw the multistage design.

But a computer test also provides the ability to adapt its design to each type of disability, at a cost that would be marginal compared to a 'paper and pencil test.'

This kind of assessment is a first in the history of French assessment programs. And like any new approach, it is subject to many future improvements.

A main question to be addressed in the future will be the test accessibility. The items used in this first adaptive test were not specifically built for students with disabilities. Because of the timeline, it was not possible, at this stage, to adapt these items to students with disabilities.

We considered an approach with assistance: each student was assisted by his/her teacher when passing the test, in the same condition than daily school-life. This assistance was defined as helping students to access to the test, and not as helping them to give correct answers. But this is not a satisfactory solution in the future.

Addressing this issue is actually addressing several different ones, as this accessibility can be improved by focusing on the content of the items, by technological changes, or by creating new innovative items to submit to students.

This assessment should be renewed every four years. This period will be used effectively to improve it, and to find answers to questions arising.

References

- D. Yan, AA von Davier and C. Lewis. 2014. Computerized multistage testing: Theory and applications. CRC Press

21. The use of images in rating scales to assess attitudes, feelings and dispositions

Christine Merrell (CEM, University of Durham, United Kingdom) and Peter Tymms (University of Durham, United Kingdom)

Across the social sciences, there are many latent variables which are of interest to researchers and others, including attitudes, feelings and dispositions to name a few. Attempts to measure these commonly involve written questions and rating scales that require Likert-type responses. There are difficulties associated with using these types of measures with young children. Many young children have limited reading skills and vocabulary levels. They may also have limited conceptual understanding of the variables being measured. As a result, self-report measures are often not feasible. Historically, to get around these difficulties, adults (parents, teachers etc.) are often asked to complete assessments on the basis of their observations of children. However, these may not accurately capture the child's own feelings; they may not choose to show their feelings through overt behaviours and the adults may not have had the opportunity to amass sufficient information. For cross-cultural and international studies, translations of written measures are problematic with subtle nuances in wording at risk of being lost. Soto et al. (2011) have suggested that self-report questionnaires are considered reliable for individuals aged 10 years onwards. For children below the age of 10 either within a country or between countries, current assessment methods are problematic. A new approach to assessing these variables is needed.

One way forward is to make use of images and to ask children to pick out someone 'who is like you' from a picture. The images can be carefully constructed to probe the construct of interest. This format of assessment can incorporate animations to more effectively convey a facial expression or a behaviour.

This paper describes the early development of a measure which uses images to assess the Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism). These traits have been identified by the OECD (2015) as being associated with positive outcomes in life and important to measure. It is vital to have an assessment of them which can be reliably used with young children from different cultures and backgrounds.

Items were created using images to represent aspects of the big five personality domains. These involved a small number of cartoon characters reacting differently to scenarios. For example one animated image consisted of a classroom scene with a teacher working on a blackboard, three small figures working at desks. One of the pupils was concentrating on their work, looking at the teacher and then back at their work; another was looking at its friend but then back at its work again; a third was watching a fly moving around the classroom and not concentrating on work at all. The behaviour of these characters was related to conscientiousness. An on-line assessment was built within which to trial the items. They were piloted on a sample of 48 participants aged 5 years – adult. The paper reports the approach to developing the items, including consideration of the types of images that would be appropriate for children of varying ages and backgrounds, a demonstration of the items themselves, and findings from the pilot study.

Overall, this novel approach to assessing latent variables appeared to be promising for children under the age of 10 years, for whom existing methods of assessment are problematic. Ideas for further development, which build upon this initial research, are discussed.

References

- OECD (2015) Skills for Social Progress: The Power of Social and Emotional Skills. http://dx.doi.org/10.1787/9789264226159-en Soto et al. (2011) Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. Journal of Personality and Social Psychology, 100 (2) 330 – 348.

Session H: Changes in assessment systems

22. Recovery from reform: The 'Sawtooth Effect' in UK secondary school assessments

Michelle Meadows and Benjamin Cuff (Ofqual, United Kingdom)

The 'Sawtooth Effect' is a pattern of performance change caused by assessment reform. Specifically, performance on high stakes assessments is often adversely affected when that assessment undergoes reform, followed by a period of improving performance over time as students and teachers gain familiarity with the new test. Several behaviours have been identified that may drive this pattern (e.g., see Koretz, 2005), such as teachers reallocating their teaching time to match the focus of the test, and successive cohorts of students being able to make greater use of past papers and mark schemes as more become available. Each of these methods serves to drive test-specific performance gains over time, whilst not necessarily increasing a students’ overall mastery of the subject.

Although some evidence for this effect can be found within the US literature (e.g., Koretz, Linn, Dunbar, & Shepard, 1991; Linn, Graue, & Sanders, 1990), evidence is lacking for the UK, and we lack estimates of the duration and size of this effect. Given that secondary school assessments (GCSEs and GCEs) are currently undergoing reform in the UK, such evidence is needed to enhance our ability to predict how students' performance may be affected in the coming years. The purpose of the current research was to gather such evidence.

As outcomes are maintained against prior attainment over time in the UK via the setting of grade boundaries (known as the ‘comparable outcomes’ approach to awarding), we were unable to investigate changes in outcomes directly. However, we were able to use a change in grade boundaries as a proxy measure for performance change, as any test-specific performance gains would have been suppressed through rising grade boundaries over time.

Raw unit grade boundaries were gathered for the just ended specification period for 1,100 GCE units and 550 GCSE units, and averages were plotted over time. A general trend was observed at both levels of study, whereby grade boundaries increased relatively rapidly over the first three years of the new assessments, changing less rapidly for the remainder of the lifespan. Additional checks ruled out alternative explanations for this pattern, supporting the proposal that performance had changed. However, when estimates of outcome change were calculated (using simulated outcome distributions), these suggested that the size of this effect was small, with estimated outcomes changing by just 2% each year for the first three years, and then by 0.5% per year thereafter.

In conclusion, the results suggest that it takes roughly three years for students and teachers to become familiar with the style and content of new assessments, meaning that we can have greater confidence that any improvements in performance after this time are due to meaningful gains in that subject area, rather than just test familiarity. Although the size of these changes appears to small, relevant parties should nevertheless be mindful of making comparisons across cohorts in the early years of a new assessment, to avoid drawing unfair conclusions about a cohort's performance simply because they were the first group of students to be entered.

These findings offer a novel contribution to our understanding of how quickly, and by how much, students and teachers are able to recover from education assessment reforms, and can be used to better predict changes in student performance following any future reforms. Plans for follow-up research focusing directly on candidate performance over time will also be discussed.

23. High stakes testing and social responsibility: an investigation into the balance of high and low order skills in high-stakes tests in England and Wales

Martin Walker (CEM, University of Durham, United Kingdom) and Kate Crabtree (Qualifications Wales, United Kingdom)

In a high stakes testing environment there can be pressure on assessment systems to focus on those things which are easily assessed as opposed to those things which need to be assessed (Au, 2007). For a high stakes testing system to be said to be socially responsible, it should probably be the case that the knowledge and skills that are assessed are those which society tends to value.

Hattie's claim (Hattie, 2014) that we privilege surface learning in assessments adds a second dimension to the issue of social responsibility for those involved in high stakes testing systems.

The pressures of high-stakes testing combined with a general tendency to test mainly simple, surface features of learning could produce a distorted picture of student achievement. It is possible that high-stakes tests could focus on lower order skills and that students who achieve high raw scores in such tests might have done little other than present large amounts of relatively simple material.

Qualifications Wales is the independent regulator of qualifications and the qualification system in Wales, sponsored by the Welsh Government. It is responsible for ensuring that qualifications and the qualifications system are effective in meeting the reasonable needs of learners in Wales, and for promoting public confidence in these. This paper will present new empirical research into the balance of surface learning (low order skills) and deep learning (high order skills) that has been found to be present to date in the most commonly taken high-stakes tests for 16 year old students in Wales. The research involves GCSE (General Certificate of Secondary Education) examinations which are high-stakes tests taken annually by the entire cohort of 16 year old pupils in Wales. Up to and including 2016, the same tests have also been taken by students in England, affecting a large cohort of approximately 600,000 pupils across England and Wales.

The outcomes of GCSE examinations shape the lives of the young people who take the examinations; these outcomes also shape the educational landscape of the countries that make use of the examination results. Whilst the proportions of students awarded the higher GCSE grades have tended to rise over the last 20 years, other evidence from studies such as TIMMS, PIRLS, PISA and the OECD Survey of Adult Skills (OECD.) suggests that the skills base in England and Wales has not improved and may even have declined.

The paper will consider the extent to which it can be said to be socially responsible to issue nationally important qualifications which appear at face value to reward high ability in students whilst the skills and knowledge required to achieve high scores on the tests might be of a lower order.

References

- Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. Educational Researcher, 36(5), 258-267.
- Hattie, J. (2014). What has 30 years of evidence in education shown us?
- Paper presented at the 30 Years of Evidence in Education, Queen Elizabeth 2nd Conference Centre, London. http://www.cem.org/30-years-of-evidence-in-education
- OECD. OECD Skills Outlook 2013. Retrieved from /content/book/9789264204256-en
- http://dx.doi.org/10.1787/9789264204256-en

24. Social and political underpinnings of educational assessment: Admission to medical schools

Avital Moshinsky and Naomi Gafni (NITE, Israel)

The number of applicants to medical schools in Israel far exceeds the number of places available; therefore, the selection procedure is high stakes and receives considerable attention from the public. Until 2004, admission to medical school was based primarily on cognitive variables: high school matriculation and the score on a standardized test (PET). Non-cognitive variables were measured by means of a personal interview.

This presentation focuses on the political and social factors that led to a revolution in the process of admission to medical schools, and on the changes that followed.

In 2002, the Israeli parliament decided to cancel the use of PET as a compulsory measure in the admissions process to institutes of higher education. As a result, the medical schools rethought their selection procedures. They used this opportunity to strengthen the role of non-cognitive variables in the process by introducing three new tools: eight multiple mini interviews (MMI), a personal biographical questionnaire and a decision-making questionnaire. Although the parliament's decision was canceled after one year, the decision to change the admission process remained valid, and in 2004 it was implemented for the first time (Ziv, et. al., 2008). It should be noted that this change was also affected by a similar trend that occurred in other places around the world (Reiter, et al., 2007).

In the new admissions process (as well as in the old one), political and social considerations have always played a role. Some of these considerations are:

Unique profiles Medical schools want to present and maintain their own unique profile, which includes having their own 'special' admissions process. This desire precludes the possibility of having one standard admissions process for all medical schools, and requires candidates to undergo more than one admissions procedure, thus increasing the cost of application in terms of time and money.

Cost–The cost of the new admissions process is high. Initially, candidates applying to more than one school had to be tested in different admission systems and pay for each of them. These high fees were perceived as unfair by the public and, as a result, the medical schools decided in 2013 to cooperate and mutually recognize each other’s admission system. This in turn resulted in a more standard, cheaper process.

Rotation of the decision makers in medical schools–every few years new deans and admissions committee chairs are appointed. Naturally, each one of them holds their own attitudes and beliefs regarding the quality of the candidates who should be admitted. As a result, there is constant rethinking regarding the admissions process. However, the new decision makers are often not aware of the history behind the process and the various considerations that have led to the adoption of the current process.

Logistics – the admissions process is complex and requires substantial resources and many raters, usually doctors who do not always have time to participate. These factors may harm the standardization of the exam and its reliability.

In summary, a variety of considerations affect the admission process to medical schools; political, social, and economic. It is important that assessment professionals make decision makers aware of these considerations and enlighten them about the possible consequences and price of each decision.

References

- Reiter, H. I., Eva, K. W., Rosenfeld, J., & Norman, G. R. (2007). Multiple mini-interviews predict clerkship and licensing examination performance. Medical Education, 41(4), 378-384.
- Ziv, A., Rubin O., Moshinsky A., Gafni N., Kotler M., Dagan Y., Lichtenberg D., Mekori Y. & Mittelman M. (2008). MOR: a simulation-based assessment center for evaluating the personal and interpersonal qualities of medical school candidates. Medical Education, 42: 991-998.

Session I: Assessment – reforms and innovations

25. Quality evidence of initial teacher education programmes: Aligning standards and graduate teachers' experiences in ever-changing social and political arenas

Claire Wyatt-Smith and Anna Du Plessis (Learning Sciences Institute Australia, Australia)

The international focus on the effectiveness of initial teacher education (ITE) and how it influences student learning poses a worldwide challenge. Recent teacher education reviews attest to the high policy and practice concerns regarding teacher preparation (Furlong, 2015). The purpose of this paper is twofold: first it presents a corpus of literature including policy reviews and completed empirical doctoral research into the preparedness of beginning teachers, with a particular focus on assessment; second, it presents information about an innovative Graduate Teacher Performance Assessment Task. The current global focus on the need for evidence of quality in initial teacher education programmes necessitates a critical analysis of how ITE programmes address beginning teachers' needs as experienced in the field. The paper provides a transnational lens with a global perspective on the interrelationship between beginning teachers' experiences and ITE programme effectiveness in preparing pre-service teachers as 'classroom-ready'. The need to provide quality ITE has resonated in evidence-based research. Goldhaber and Walch (2014) and Coughlan (2014) noted that the academic outcomes students obtain relate to high-quality teachers. Hence, it is a concern that school leaders perceive beginning teachers as unprepared – not ready for what awaits them in classrooms, and in particular, in classroom assessment.

Global research about quality ITE and assessment evidence has provided a wide range of assessment instruments, for example, Teacher Performance Assessment (edTPA), Performance Assessment for California Teachers (PACT), and the Australian Graduate Teacher Performance Assessment (AGTPA) presented in this paper. The development of the AGTPA instrument is deeply embedded in Vygotsky's (1978) social constructivist learning theory, assessment as social practice (Broadfoot, 2002; Broadfoot & Black, 2004), theorising of standards (Sadler, 1989) and research into professional standards and assessment work of teachers (Wyatt-Smith & Looney, 2016). The development of the AGTPA is also informed by two main studies: the first is an investigation of all Initial Teacher Education Programs in Queensland, Australia. Programs were considered against the requirements of Australian Professional Standards for Teachers (Australian Institute for Teaching and School Leadership, AITSL, 2011), and National Program Standards for the Accreditation of Initial Teacher Education Programs in Australia (AITSL, 2015). The second study is Du Plessis' doctoral research that included semi-structured interviews, document analysis and classroom observations with a particular focus on enacted classroom assessment. The study exposed gaps in the preparedness of beginning teachers for classroom assessment. The analysis of this corpus (published literature; government reports; empirical research; ITE program analyses) opens the space for addressing a key question: 'What evidence could meet the tests of validity and reliability to show graduate readiness for teaching and assessment' The search for answers calls for explicit valuing of human capital, consistent with Gadamer's (1975, 1976) hermeneutics philosophy as its theoretical framework to look deeper into the perceptions and life-worlds of beginning teachers in order to reveal assessment insights not otherwise available. Additionally, Heidegger's (1962) theories of 'being in the world' (p. 174) encourages an understanding of the need for 'belongingness' and 'at homeness' of graduate teachers in their first professional placements. The paper's argument that ITE and quality evidence programs should set graduate teachers up for success in teaching and assessment practice is embedded in an in-depth understanding of real-life experiences. We argue that human capital should not be overlooked in the design, implementation and moderation of authentic Graduate Teacher Performance Assessment Tasks. The AGTPA task is presented with suggestions for further research on the benefits of aligning beginning teachers' lived experiences with ITE programmes, policies, recruitment and placements.

26. Preparing for College Success: Exploring the Impact of the High School Cambridge Acceleration Programme on US University Students

Magda Werno and Stuart Shaw (Cambridge Assessment, United Kingdom)

This study investigated the impact of the Cambridge international acceleration programme based on participants' perceptions of the effectiveness of the programme in helping students in their transition to college-level study. The programme includes the International Advanced Subsidiary Level (AS Level) and the International Advanced Level (A level) – qualifications offered by schools for 16–19 year olds.

High stakes acceleration programmes are instrumental in shaping educational goals and processes. Their impact can have important implications for teaching and learning, as well as on other stakeholders outside the classroom. It is important, therefore, that any impact study takes account of the perceptions of its stakeholders, because their attitudes towards the programmes may be relevant to its validity.

In this research project, 104 students across all year groups in one American university (Florida State University) responded to an online questionnaire designed to elicit their attitudes, perceptions, and subjective experiences associated with the programme. The study aimed to provide an important insight into the attitudinal and subjective aspects of the transition from Higher Education (HE) to college, to enrich the previously reported quantitative research undertaken at the same university (Shaw & Bailey, 2011; Shaw, Warren & Gill, 2014).

The Cambridge acceleration programme was described as challenging and academically rigorous, allowing students to develop a range of skills and learning attitudes which subsequently helped them adjust to the demands of university (Conley, 2015). These included time management, writing skills, critical thinking, the ability to work under pressure, as well as independent study and perseverance. A large proportion of respondents highlighted similarities between their high school and college experiences in terms of the workload, subject knowledge and understanding required, and study techniques. As a consequence of completing the programme, the students developed skills perceived as relevant to the demands of university, which facilitated the transition from high school to univeristy. Some students also suggested that certain aspects of the first year college curriculum appeared less demanding compared with the Cambridge programme, although these perceptions varied.

Conversely, the majority of students reported finding at least some aspects of the transition difficult. These included the increased amount of coursework, limited support from college tutors, the necessity to study independently, and problems with concentration and motivation. These comments could reflect individual differences in academic abilities and engagement, as well as the quality of teaching at high school. Importantly, students' perceptions of the Cambridge programme and their initial experiences at college, as well as the success of the college-level programmes run in high schools could depend on the preparedness of both students – to cope with the increased level of demand and pressure, and teachers – to deliver such advanced programmes.

The findings of this study provided important insights into the impact of the Cambridge acceleration programme on college readiness and students' transition to HE. The programme appeared to help participants develop a range of skills which were perceived as important in the context of university study. The implications for future research and the provision of the Cambridge acceleration programme within the US context are discussed.

References

- Conley, D. (2015). College-Readiness Assessments: Be Careful What You Wish For. Education week, Opinion, Learning deeply.
- Shaw, S. D. & Bailey, C. (2011). Success in the US: Are Cambridge International Assessments Good Preparation for University Study? Journal of College Admission, No.213, pp.6-16, Fall
- Shaw, S. D., Warren, J. and Gill. T. (2014). Assessing the impact of the Cambridge international acceleration program on US university determinants of success: a multi-level modelling approach. College and University Educating the Modern Higher Education Administration Professional. Vol. 89, No.4 (Summer 2014).

27. Snapshots of deep learning over time: A novel approach to measuring student progress

Ian Jones (Loughborough University, United Kingdom) and Brian Henderson (No More Marking Ltd., United Kingdom)

Traditional academic subjects such as mathematics and language are increasingly viewed as involving complex, ill-defined abilities such as creativity, sustained reasoning and communication skills (Suto, 2013). Such abilities can be evidenced using open-ended tests that are designed to prompt as varied and unpredictable a range of student responses as possible (e.g. Bisson, Gilmore, Inglis & Jones, 2016). An example of such a test in mathematics is the question prompt 'What are the differences between the mean, the mode and the median? Give examples of when they are appropriate for summarising data.' However, it is difficult to design corresponding scoring rubrics, which traditionally assume predictable and uniform responses in order that a population of scorers can apply points objectively. Recent developments in comparative judgement (Pollitt, 2012) have enabled this barrier to be overcome by doing away with rubrics and instead collating experts' pairwise judgements of student responses to produce reliable outcomes.

In this presentation we will report the results of using a comparative judgement approach to measuring student progress in mathematics and English. An initial pilot study with students aged 11 and 12 years in seven secondary schools demonstrated that the procedure yielded high estimates of reliability and validity in both subjects tested. The pilot study also enabled us to scrutinise the student responses to, and analyse the psychometric performance of, the specially-designed and open-ended test questions. The subsequent larger study with students of the same age in about 75 schools demonstrated the robustness of the approach for measuring student progress over the course of an academic year. Again, the performance of individual test questions was monitored by scrutinising responses and psychometric techniques. This activity enables the selection and refinement of test questions, improving further the robustness of outcomes in future work.

Three components were critical to the success of the studies. First, a robust and usable online comparative judgement engine. Second, the recruitment of judges who are experts in their fields and undertake their assessments sincerely and thoughtfully. Third, the design of test questions that are truly open-ended, and that are perceived to have high face validity by end users. Indeed, the tests proved popular with many teachers who reported that their students' responses provided valuable, intuitive snapshots on deep learning, as compared to more traditional test formats that are designed with rubrics and scoring in mind. In these nascent days of applying comparative judgement to educational assessment the research focus to date has necessarily been on technology and logistics. The design of test questions specifically for use with pairwise judging has received less attention but is now of equal importance.

In conclusion, we will argue that the approach enables robust and meaningful progress data to be generated for teachers and schools, while privileging and promoting the highly-valued complex abilities that have until recently remained elusive in assessments.

References

- Bisson, M., Gilmore, C., Inglis, M. & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. International Journal of Research in Undergraduate Mathematics Education. Online first.
- Pollitt, A. (2012). The method of Adaptive Comparative Judgement. Assessment in Education: Principles, Policy & Practice, 19, 281–300.
- Suto, I. (2013). 21st Century skills: Ancient, ubiquitous, enigmatic? Research Matters, 15, 2–8.

Session J: The social effects of assessment results

28. Unleashing the power of human capital: Workforce assessment, strategy, skills and standards

Daniela Muresan (ETS Global, The Netherlands)

ETS Global will share its expertise in global assessments to discuss what it means to be work ready in a global context, the skills necessary to develop a competitive workforce, and how to define useful common standards and metrics.

In recognition of the challenges of building the right skills and turning them into better jobs & lives, the OECD has launched a Skills Strategy Action Plan to promote global awareness and coordination. OECD is calling for a skill agenda that addresses the mismatch between worker skills and employer needs – an evolution of workplace expectations due to technology and global competition as well as efforts to improve both efficiency and quality.

ETS has developed WorkFORCE® Assessment for Job Fit, an on-line, 20-25 minute adaptive assessment which enables organizations to evaluate the overall job fit of candidates, as well as measure six behavioural competencies associated with critical drivers of job success: initiative & perseverance, responsibility, teamwork & citizenship, flexibility & resilience, problem solving & ingenuity, customer service orientation. The personality attributes are measured using a 104-item pairwise preference test that is administered with the FACETS™ engine, a computerized adaptive testing (CAT) environment.The six behavioural competencies are tailored to the type of job that an organization is recruiting for in the local cultural and workforce environment and the behavioural skills needed to perform successfully in that job. The test is designed to help employers efficiently and effectively uncover the 'right' candidate (for screening and selection decisions) and/or employee (for development, career promotion decisions).

Employers need an increasingly educated workforce, not only in the foundational level of academic competencies and cognitive skills, but in 21st Century Skills. The goal is to link a set of measures developed and validated in an international context to standards and benchmarks of workplace success. Foundational cognitive, workplace communication and behavioural skill components are essential for workplace success. As such, they are the platform upon which job-specific knowledge, skills and abilities rest.

Topics to be covered as part of the session include:
- What does it mean to be 'work ready?'
- How to define common standards and metrics in ways that are meaningful to employers
- How to build skills that are transportable

In conclusion, establishing work readiness benchmarks and standards within a multi-national context will accelerate adoption and use of employability standards essential to implementing skills, policies, employer practices and education system commitment to address essential work readiness standards required of work-ready individuals.

References

- Naemi B., Burrus J., Kyllonen P., Roberts R.D., (2012) Building a Case to Develop Noncognitive Assessment Products and Services Targeting Workforce
- Readiness at ETS
- OECD (2012). Better skills, better jobs, better lives: A strategic approach to skills policies
- Austin, J. T., Mellow, G. O., Rosin, M., & Seltzer, M. (2012). Portable, stackable credentials: A new education model for industry-specific career pathways.
- McKinsey Global Institute (2012). Help wanted: The future of work in advanced economies.

29. Supporting stakeholder trust in A level Modern Foreign Language outcomes: is there a native speaker effect?

Rachel Taylor (Ofqual, United Kingdom)

A levels are high stakes assessments typically taken at age 18 in England, Wales and Northern Ireland. The results that students achieve are put to many uses: they allow students access to University level education and can open the door to further employment opportunities. This highlights the importance of ensuring fairness in the results that students achieve, and the responsibilities that awarding organisations and regulators face in ensuring that students are awarded the grades that they deserve.

This paper focuses on A level Modern Foreign Languages (MFL), qualifications that were designed for students that are acquiring the MFL as a second language (i.e. non-native speakers). The paper discusses the findings of a study that explored the potential effect of examination entries from native speakers on overall national results in A level French, German, Spanish, Russian and Italian. The research was planned in response to stakeholders' concerns that the proportion of native speakers sitting A level MFL is increasing, and that students for whom the MFL is a second language are being disadvantaged because of this. Such concerns, even if unfounded, have fuelled a lack of trust in A level MFL grades from key stakeholders, including prominent language and teaching associations. This has been coupled with a steady decline in the uptake of some of the largest entry A level MFL subjects (JCQ, 2015), prompting concerns of a link between the two.

Currently, there is little evidence in the literature to support or refute stakeholder concerns in relation to the effects of native speakers in A level MFL. The aim of this study were therefore threefold: i) to quantify the scale of native speakers certificating in A level MFL; ii) to consider the potential effects of native speaker characteristics on the results that students achieve; and iii) to address stakeholder concerns around the potential effects of native speakers on A level MFL results.

The research focused on A level French, German, Spanish, Italian and Russian. Schools and colleges with students entering A levels in these subjects with any of the five UK awarding bodies in summer 2016 were contacted to take part in the research and data were collected from schools about students' native speaker characteristics. Teachers were asked to indicate, to the best of their knowledge, which students were native speakers in each subject, and students were asked to complete a questionnaire about their language expertise and proficiency. This was an adapted version of the LEAP-Q, a questionnaire that has been validated and used in a number of research studies (e.g. Marion, Blumenfeld, & Kaushanskaya, 2007). The information provided by students and teachers was used to consider students on a continuum of native to non-native speaker, and to explore any associations between native speaker characteristics and the grades that students achieve.

The results focus on considering the prevalence of native speakers in A level MFL, how native and non-native speakers perform, and the potential implications for the grades that students achieve. The evidence gathered through this study will also be used to consider the potential implications of native speakers on standard setting and ensuring that students achieve the grades that they deserve. In doing so, the research will seek to address stakeholder concerns and build trust in high stakes assessment outcomes.

References

- Marian, V., Blumenfeld, H.K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. Journal of Speech Language and Hearing Research, 50 (4), 940-967.
- JCQ (2015). Appendix–GCE Trends 2015. Retrieved from http://www.jcq.org.uk/examination-results/a-levels/2015/gce-trends-2015

30. PISA Results as a Mirror of Social Stratification Reproduction: A Story from Serbia

Jelena Radišić (University of Oslo, Norway), Aleksandar Baucal (University of Belgrade, Serbia) and Jasminka Čekić Marković (Center for Education Policy, Serbia)

Each cycle of the Programme for International Student Assessment (PISA) offers participating countries an insight into quality, efficacy and equity of own education system. Result in Serbia show relatively high level of equity within the system, meaning the quality of the system is equal for all students irrespective of their socio-economic background (SES). However when observing effects of SES on student achievement one should not neglect its secondary effects on student enrolment into different education programmes. This means that due to the lack of important cultural and educational resources students of lower SES may obtain lower achievements or may not be able to enrol into specific education programmes despite own high attainment (Bourdieu & Passeron, 1977/1990; Becker, 2003; Sirin, 2005). In Serbia such examples may be found when transiting from primary school programmes to upper secondary education. Around 25% of students are enrolled in grammar schools, with programmes preparing them for higher education, and 75% of students opt for vocational education (VET, tracks lasting 3 or 4 years). Secondary SES effect in this case would be demonstrated if lower SES successful students would have less chance to enrol in grammar schools in comparison to the other successful students of different SES. Thus we investigate what the odds are for students of highest educational achievement, but different SES background, to be enrolled in grammar schools, which are considered to be of higher quality and better prepare students for higher education?

Method

Data from PISA 2006, 2009 and 2012 were used in the current analysis. SES was defined as a composite measure of parents' education status, parents' occupation, occupational prestige, economical status and cultural resources the family may possess (OECD, 2013). The analysis was performed taking into account educational streaming (grammar school or VET), SES and achievement. 20% of students that achieved highest scores in the domains of mathematical, reading and science literacy were secluded for the analysis from the sample in each of the cycles (2006, 2009, 2012). Following, the selected 20% of students were differentiated according to the SES quintile they belong to. Assuming that these students are competent enough to succeed in the grammar program and higher education after that, then it is expected that successful students have the same chance to be enrolled in the grammar school regardless of SES. In order to check this assumption enrolment probability in grammar schools is calculated for students from each SES quintile.

Results

The analysis shows that most successful students in Serbia belonging to the lowest SES have about three time lees chance to be enrolled in grammar schools, as an opposite to other students belonging to the most successful group. At the same time students from the highest SES group have two times more chances to be enrolled into the same track. The results are consistent for all three cycles. Evidences also show that while the percentage of students with the highest SES who are enrolled in the grammar school is between 17% and 20% points higher than the average, this number is in a slight decline for the most disadvantaged students. While in 2006 the percentage of the most successful students with the lowest SES, enrolled into grammar school was 31% points lower than the average for the entire group of the most successful students, in 2009 it decreased to 27 percentage points lower and 25 percentage points lower in 2012.

The results will be discussed in the light of the state's higher education entrance examination and students' scholarships and loans system, and data from other Balkan countries with equal education streaming.

Session K: The micro-politics of assessment

31. The complex interplays between assessment and learning that shape writing development during the transition to university from A-level

Natalie Usher (OUCEA, United Kingdom)

In many societies, students take high-stakes tests with social and political purposes throughout their educational careers. While policy and practice has come to emphasise the importance of formative assessment in the classroom, high-stakes assessment continues to shape students' experience of learning, even after testing. The present study addresses the complex relationship between assessment, learning and writing during the transition to university in England. In the first year of university, students are often required to rethink the strategies they used to prepare for A-level writing.

The research investigated the impact of peer assessment on students' writing development in an exam essay with different conventions and scope to both A-level and other degree writing. The target genre was a timed English literature essay examined at the end of the first year and determining progression. The first-year participants saw this exam essay as offering more 'freedom' and choice than A-level. During the intervention, 20 students took part in a series of writing workshops using comment-only peer assessment. Having assessed and discussed four examples, participants wrote a practice essay then exchanged anonymous written comments, using a holistic approach. Following Sadler (2010), the criteria were not pre-determined, but instead negotiated after assessing three examples. The theory of action drew on the Winne and Hadwin (1998) model of self-regulated learning (SRL). Learners self-regulate when they proactively monitor and adapt their strategies for learning where necessary. Applying SRL together with a case study approach allowed the influence of the peer assessment intervention and prior experiences to be traced. A range of qualitative data were gathered to evaluate learning, including writing samples, peer assessors' comments, and reflections. Ten case study students also took part in a pre- and post-intervention writing session, involving a think-aloud protocol, timed essay and interview.

The presentation will use case study data to highlight the complex interaction between influences on participants' writing development, including prior experiences, such as A-level. Through peer assessment, participants developed their declarative knowledge of what makes a good quality essay, as evidenced by inductive analysis of annotations and peer reviews. Visualising writers' development over the course of the study highlights shifts in the way many participants self-regulated the writing process. For instance, task perceptions became more sophisticated, and writers applied new strategies for processes such as planning. Writers' beliefs about writing also evolved, with the differences between A-level and degree becoming more apparent, and participants beginning to see an alignment between longer formative essays and the exam essay. However, a few writers took a 'risk management' approach and adapted much less. They chose to use revision strategies developed during A-level preparation which allowed them to memorise and reproduce knowledge, but not necessarily successfully reshape it in response to the more complex university questions.

While the essay at the focus of this research is not high-stakes on the national stage, participants saw it as carrying high-stakes on a personal level. The examinations do not contribute to final degree outcomes, but do determine progression. In the words of one participant, the exams test 'all the sorts of reasons... you should have been let in here', justifying admission as well as progression. The data show how affective responses to assessment shaped whether writers took a 'risk management' or a more proactive approach. Thus, the findings highlight that even where the focus of research is on intra-individual learning, the social situation of assessment comes to the fore and shapes the decisions and choices students make.

32. The micro-politics of the school in the process of changing assessment cultures

María Teresa Flórez Petour (University of Chile, Chile)

There is wide consensus in the literature on Assessment for Learning (AfL) about the positive effects of this approach as perceived by teachers, students and the school community (Tierney and Charland, 2007; Hodgson and Pyle, 2010; Leahy & Wiliam, 2012; Flórez and Sammons, 2013; Flórez, 2015). AfL is often characterised in research as generating changes in classroom relationships and practices that are consistent with socio-constructivist perspectives on teaching and learning (Stobart, 2010). However, AfL is seldom found in practice and evidence around the obstacles it faces in its process of enactment in schools is abundant (Hopfenbeck et al, 2015). Lack of commitment from senior staff and differences in beliefs around assessment in the school community are mentioned among these obstacles (Hopfenbeck et al, 2015). AfL, as an innovative repertoire, has to struggle with long-standing cultures of assessment whose values are not consistent with the new approach (Flórez, 2015), thus generating struggle and resistance in the process of its implementation. Studying AfL from the perspective of the conflicts, tensions and synergies it generates among the actors of the school community during its process of enactment becomes, therefore, crucial.

Stephen Ball’s conceptualisation of the micro-politics of the school aims at overcoming traditional organisational approaches, where the school is understood from a priori models based on ideas of consensus, shared goals, authority, motivation and ideological neutrality (Ball, 2012). Unlike these approaches, Ball proposes studying schools with no a priori lenses and as spaces of power struggle, characterised by diversity of goals and ideologies, conflicts of interest among groups and political activity (Ball, 2012). Following this conceptualisation, the paper addresses the tensions and facilitating aspects in the process of changing the assessment culture of a school through a professional development programme around AfL. The study on which this paper draws chose two highly selective state schools as cases for research, due to their higher potential for resistance to change. Their reliance on traditional testing regimes as a self-legitimating tool for their identity of success and prestige makes them interesting cases for research.

Drawing on materials derived from the AfL professional development programme under scrutiny (essays, session logs, evidence of assessment practices, design of assessment criteria, among others) as well as on semi-structured interviews, focus groups and workshops with different actors of the school community, the study reconstructs the map of relationships inside the school in connection to the assessment discourses that circulate between them.

References

- Ball, Stephen (2012). The Micro-Politics of the School. Towards a Theory of School Organization. London: Routledge.
- Flórez, T. and Sammons, P. (2013). A literature review of Assessment for Learning: effects and impact. Oxford: University of Oxford, Department of Education and CfBT.
- Flórez, T. (2015). Systems, ideologies and history: a three-dimensional absence in the study of assessment reform processes. Assessment in Education: Principles, Policy & Practice, 22(1), 3-26.
- Hodgson, C. and Pyle, K. (2010). A literature review of assessment for learning in science. Slough: UK National Foundation for Educational Research.
- Hopfenbeck, T; Flórez Petour, M.T.; Tolo, A. (2015). Balancing tensions in educational policy reforms: large-scale implementation of Assessment for Learning in Norway. Assessment in Education: Principles, Policy & Practice, 22(1), 44-60.
- Leahy, S., & Wiliam, D. (2012). From teachers to schools: Scaling up professional development for formative assessment. In J. Gardner (Ed.), Assessment and learning (2nd ed., pp. 49–72). London: Sage.
- Stobart, G. (2010). Tiempos de pruebas: los usos y abusos de la evaluación. Madrid: Morata.

- Tierney, R. and Charland, J. (2007). Stocks and Prospects: Research on Formative Assessment in Secondary Classrooms. Online Submission, Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 11.

33. Monitoring of student achievements in Mathematics as an effective instrument to adjust individual learning paths for students and to enhance didactical tools of teachers

Laila Issayeva, Daniyar Temirtassov, Baimurat Akhmetov, Yerbol Nurguzhin (Nazarbayev Intellectual Schools, Kazakhstan), Nico Dieteren and Frans Kleintjes (Cito, The Netherlands)

Having entered the top 50 most competitive countries among 148 countries in the World Economic Forum's Global Competitiveness Index in 2013-2014 (OECD, 2014), Kazakhstan set a new goal to progress even further to become one of the top thirty most developed countries in the world. President Nazarbayev (2012) claimed that, that the country has 'all possibilities: resources, educated people, and a united nation to reach that goal' (Strategy: Kazakhstan 2050). Due to the purpose for the prompt improvement of its economic situation, Kazakhstan aims to strengthen the role of education in development of the competitive human capital by implementing progressive reforms based on the local needs and oriented to the international standards. President Nazarbayev (2011) emphasized the necessity of forming intellectually, physically and spiritually developed citizens of the Republic of Kazakhstan in order to satisfy national needs in education that can provide success in the rapidly changing world (SPED RK for 2011-2020, 2010).

Such ambitious goals became realizable with the support of the establishment of Nazarbayev Intellectual schools (hereinafter – NIS) – a flagship of Kazakhstani secondary education system. The implementation of the new subject programmes, in particular in Mathematics, which were developed by NIS collaboratively with Cambridge University, started in 2012. Along with that, criteria-based assessment system was created aiming to provide transparent and fair assessment of student academic performance, to increase student motivation for developing skills necessary to achieve the expected learning outcomes, etc. These innovations required the creation of an efficacious assessment tool that could provide reliable data on student achievements in dynamics, and, therefore, on the quality of the innovative subject programme. To address this crucial need NIS decided to develop monitoring of student achievements in Mathematics.

This monitoring which has been developed by NIS together with the Institute of Educational Measurement Cito, the Netherlands, since 2012, aims to diagnose student performance and to help the progress of an individual learner. Candidates to study at NIS go through a strict selection test checking their abilities and competencies in Mathematics and languages. Therefore, NIS accept only talented and gifted students who are supposed to have high achievements in learning. However, differential monitoring results show that the distribution of NIS students across student achievements obeys a Gaussian curve. The Normal distribution curve was explained in terms of education by the American scholar as following: a large number of students do moderately well; some do worse than average, and some do better; and a very small number get very high or very low results (Muhammad, 2011).

Therefore, monitoring reports benefit both main stakeholders involved into the studying process; teachers are informed about student knowledge levels and are empowered to provide appropriate professional support, and students, in turn, are aware of their weaknesses and strengths, and are invited to take responsibility for the further improvement in Mathematics. Thus, monitoring reporting helps to define not only students who need extra help from teachers, but also students with high performance which needs to be used for further development their talents. Monitoring results also let policy makers use gathered information for making relevant decisions in order to improve coordination of a curriculum and a subject programme (Bartle, 2007).

While data processing of the student monitoring results in some grades a high percentage of item missing was discovered. This fact facilitated the authors of this paper to conduct sociological research among potential stakeholders in order to explore the effectiveness of the monitoring. Findings from this research will add to the knowledge base in the educational realm, and contribute into further improvement of the developing monitoring system in Mathematics at NIS of Kazakhstan.

Friday 4th November

Session L: Assessment quality

34. Improving marking quality and examiner experience: zoning scripts with generic answer booklets

Matthew Glanville (International Baccalaureate, United Kingdom) and Martin Adams (RM Results, United Kingdom)

Students use generic answer booklets to respond to examinations which have a choice of questions or require extended responses. This approach means that the location of every question in the script can be different for every student. This makes the role of the examiner significantly harder because the answer to a question can be spread throughout the script. In addition, the examiner must have marking expertise in all of the different optional questions to be able to mark to the required quality.

IB and RM results have been working collaboratively to develop a solution to automate the process of zoning every script. The ‘ebookmarking’ solution uses a modified answer booklet design to optimise the capture of candidate question numbers using Intelligent Character Recognition (ICR) and manual correction of inaccurate captures. These captured question numbers are used to attribute each part of the answer booklet to a question. The result means that examiners can be allocated to mark specific questions and the e-marking software, RM Assessor automatically collates all of the responses for each particular question.

In this session, IB will elaborate on the benefits of zoning generic answer booklets to marking quality, and the utilisation and retention of examiners. RM Results will describe the journey of researching, developing and trialling an automated ‘ebookmarking’ solution.

The research and trialling phase evaluated the use of three different question recording mechanisms and then trialling two different answer booklet designs with 900 students. The completion accuracy of each method was compared both at a question level and a script level. This phase concluded that both Optical Mark Recognition solution and Intelligent Character Recognition solution were viable options. The Intelligent Character Recognition solution was defined for operational use and developed for piloting.

The solution was piloted with 12,000 students in November 2015 and we will evaluate the solution against the original business requirements: improving the quality of marking, improving examiner utilisation and improving examiner retention. To estimate the improvements to quality of marking, we will compare examiner performance on seeds using ‘ebookmarking’ components with examiner performance on seeds for components not using ‘ebookmarking’. IB use qualification scripts to ensure each examiner can mark to quality prior to marking live student scripts. For utilisation of examiners, we will compare the proportion of examiners which were used to mark at least one question for component and compare to the proportion of examiners who would have been permitted to mark the component as a whole paper. To evaluate the retention of examiners, we will use this qualification data and seeding data to estimate the proportion of retained examiners and compare these metrics with historic data. Examiner feedback will be used to estimate the retention of using 'ebookmarking' on examiners marking in future sessions.

To understand the accuracy of the overall solution, 1400 scripts were analysed in depth to determine the question level inaccuracy and the script level inaccuracy. All question numbers were captured to create a reference set of data. This reference data was used to re-zone the scripts and compare to the zoning achieved in the live session. Examiners raised exceptions to operational staff where they believed zoning errors existed. Further analysis has categorised these exceptions as zoning error, candidate error or examiner error. In addition, operational metrics were determined including the proportion of candidate scripts with unknown zones. The key findings from this analysis will be presented in the paper. Finally, we will share the next research activities for future evaluation.

35. What makes a good seeding script? Perceptions from Principal Examiners of a UK awarding body

Martina Kuvalja and Simon Child (Cambridge Assessment, United Kingdom)

One of the fundamental tenets of assessment delivery, as required by the regulator for qualifications (Ofqual,2011) in the United Kingdom, is the effective monitoring of examiners' quality of marking. 'Seeding' is one of the key marking monitoring approaches used for UK general qualifications. 'Seeding' scripts are selected and pre-marked by a Principal Examiner (PE) and are utilised to provide an on-going assessment of examiners' application of the mark scheme. Approximately one in 20 scripts which markers receive is a 'seed' from which a judgement is made on their marking performance. If conceptualised as part of an 'assessment' of examiners' marking, the use of seeding scripts must be valid in the sense that they map on to the genuine experience of the marker, and be able to discriminate between markers that are applying the mark scheme adequately, and those that are outside of pre-defined parameters. The characteristics of seeding scripts, and our understanding of the process of selecting them, are prerequisites for the future improvement of marking quality.

Whilst there has been previous research that has examining how seeding items can influence marker agreement (Black, 2010), there has been little research investigating PE's perceptions of what makes a good seeding script. This study aimed to explore the criteria used by the PEs when selecting seeding scripts. As seeding scripts are selected at a similar time to scripts used for the purpose of standardisation, a second aim of the research was to compare and contrast PE's criteria for selecting these script types.

The present study used a survey design. PEs (n = 96) were asked about the decision process in the selection of standardisation and seeding scripts. The majority of the PEs reported that the seeding scripts were generally selected on the basis of three main criteria: that they were 'straightforward to mark' (95.6%); representative of the performance of the cohort (92%); and legible (77.2%). Overall, when criteria for all units were compared, there were no significant differences between criteria used for selecting standardisation and seeding scripts. However, the majority of the PEs (60.6%) also indicated that standardisation and seeding scripts differed. PEs felt that seeding scripts should consist of 'straightforward' answers only, while the standardisation scripts can have some 'ambiguous' answers. In addition, the same group of PEs only selected clear and readable seeding scripts, while the selected standardisation scripts included some unclear responses.

The study identified that the divergent roles of standardisation and seeding scripts influence the criteria PEs adopt for their selection. Overall, PEs felt that the purpose of seeding scripts was not to 'test' the examiners, but rather to maintain the marking standard for examiners who had previously already 'passed' the test of using the mark scheme appropriately, as determined by their performance on standardisation scripts. It is then not surprising that the main difference between the seeding and standardisation scripts was related to straightforwardness in marking. It is an open question, however, as to whether this method of selecting seeding scripts reduces the validity of the monitoring process, as the scripts are not necessarily representative of the entire range of responses (at least in terms of ambiguity or perceived difficulty to mark). However, another important consideration is that seeding scripts should not be able to be identified by examiners during their normal marking activity. Therefore the selection of 'straightforward' seeding scripts may prevent their identification, thus increasing their efficacy. This study contributes to the understanding of the underpinning concepts contributing to overall quality of marking, one of the main concerns within the assessment industry.

36. Implementing the Assessment Agenda for Intermediate Vocational Education in the Netherlands: Developing Flexible Digital Exams

Cor Sluijter and Marieke van Onna (Cito, The Netherlands)

In 1996 a new Law on Adult and Vocational Education came into effect in the Netherlands. This resulted in a principal change in the nature of exams in Intermediate Vocational Education (IVE). Nationwide central exams were replaced by school exams. IVE schools became responsible for developing their own exams and ensuring their quality. The most important reason for this change was that the Dutch government wanted to ensure a maximal fit of the competencies of students with an IVE diploma to the demands of the labour market.

This principal change marks the onset of a 20 year long period in which several changes of the original exam system took place. All with the purpose of better assuring the quality of the exams in IVE and to increase stakeholder's trust in the worth of the IVE diploma's. The actual exam system can be described as a hybrid system, consisting of a mix of instruments that together guarantee that a student receiving the IVE diploma for a certain level can be considered to be a competent starting professional on this level. The mix of instruments consists of different components ranging from simple multiple choice tests for theoretical knowledge to complex performance assessments. Exam components are still partly developed by the schools themselves, for efficiency reasons for the larger part in collaboration. But each year more and more exam components are purchased from specialized testing companies.

A structural change to the exam system was implemented in the school year 2014-2015. Standardized compulsory national exam components for Dutch and arithmetic's were reintroduced for IVE level 3 and 4. These will be followed in the school year 2017-2018 by a standardized compulsory exam for English for IVE level 4. This change came about by a growing concern over the achieved level of Dutch, English and mathematics for IVE students. In 2010 the decision was made to make certain that candidates graduating for IVE levels 3 and 4 would meet certain minimal standards for these subjects. National exams were to be reintroduced instead of having IVE schools develop their own.

Cito was requested to develop exams that would optimally fit within the complex IVE exam system. In this presentation the political and educational context is described, together with the methodology used. With the help of Item Response Theory item banks were constructed making computerized exams with different equated versions possible resulting in an exam system with maximal flexibility. The presentation covers the construction, pretesting, analysis and standard setting phases of the construction process, including the equating process of all different exam versions. In its current form the system allows students to sit their exams almost throughout the whole school year, thus interfering as little as possible with the curriculum of IVE-schools. And in accordance with the partly personalized educational tracks within IVE.

With the inclusion of a national compulsory component, the IVE exam system has finally reached its optimal form. The combination of a mix of instruments being developed by the schools themselves or purchased from specialized test suppliers, supported by a small set of compulsory national exams ensures that the interests of all stakeholders are looked after and the value of IVE diploma's is guaranteed both for the labour market and for receiving schools for Higher Vocational Education. IVE graduates enter the labour market more prepared than before and have a greater chance of succeeding in schools for Higher Vocational Education.

Session M: Assessment policies: what's behind?

37. Examining the 'global revolution' of English language learning

Elizabeth Shepherd, Wahida Amin and Victoria Ainsworth (British Council, United Kingdom)

Research on the English language has been influenced by 'the present, apparently unassailable, position of English in the world', with scholars asking 'whether we can expect its status to remain unchanged during the coming decades of unprecedented social and economic global change' (Graddol 1996). Contributions to the field (Graddol 2006, 2010) (Erling 2014, 2015) continue to shape the conversation about the far-reaching impact of English language communication on economic, societal and individual human growth into the next millennium.

One significant socio-political manifestation is the phenomenon of a 'global revolution' of English language policy development, which has taken multiple national forms across all regions of the world (Enever and Moon 2010). In Mexico, for example, the Ministry of Public Education has implemented the National English Program in Basic Education from 2009, an ambitious policy for English learning from preschool to secondary school. Mexico became the first country in Latin America to include English in all thirteen years of public school curriculum.

The purpose of this presentation is firstly to examine this 'global revolution' and the consequences, by tracing key themes in language planning across regions, focusing particularly on assessment. This will be done by comparing English language learning policy, provision, impact and ability and featuring preliminary findings from a new global study analysing English in-education policy trends in 24 countries.

The pressures for English language programmes can originate in both national government policy, and individual parental concern that in a globalised world, English is required for societies and their populations to be competitive and economically prosperous. This presentation will secondly scrutinise the way in which future economic sustainability and growth is the principle motivation in education policies to up-skill national populations with English language communication skills. While these policies have been described as a global shift in language learning, much debate and discussion remains on the variety of contexts and applications of English language provision in different education systems (Enever and Moon 2010).

The English language learning system explored will also be defined by 'the broad triangle of curriculum, delivery and assessment, which together form an integrated whole without which learning is unlikely to be facilitated' (O'Dwyer, O'Sullivan 2011). These three elements will be considered to be defined and driven by national policy. Also important to explore will be the opportunity to learn within the curricula that can promote learning for all students, such as learning material, facilities and teachers (Stevens 1993), the relational motivations and backgrounds of these participants, and student learning outcomes.

We argue that a multi-faceted approach is needed to understand this 'global revolution' of English language learning in response to the views of language testing theorists that language tests act as agents of cultural, social, political, educational and ideological agendas. Within this branch of language testing theory, practices of language testing are also linked to analysis of language teaching performance. It's commonly believed, however, that the knowledge created via a test is 'narrow and simplistic [...] it is mono-logic based on one instrument which is used on one occasion, detached from a meaningful context.' Shohamy (1998) goes on to explain that using tests can provide 'a quick fix', and an instant solution, however analysis of data captured via this method alone overlooks the complexities of broader subject matter and is un-meaningful for repair.

38. Educational governance and PISA: challenges and prospects in the Cypriot context

Myria Vassiliou, Aristotelis Zmas (European University of Cyprus, Cyprus) and Michalis Michaelides (University of Cyprus, Cyprus)

During the last decade the PISA study has gained a prominent position in discussions on the formation of educational policies. Many scholars regard PISA as a 'worldwide engine' which assesses, classifies and allocates students, teachers and educational systems (Meyer & Benavot, 2013, p. 9). Political actors use the PISA findings either for educational reforms (Lawn & Grek, 2012) or in order to support and promote policies that have already been in place or changes that have already been introduced (Pons, 2012). In many cases, countries introduced educational reforms after the announcement of the results, such as Germany (Ertl, 2006) and Japan (Takayama, 2012).

The PISA study has been quite controversial. On the one hand, some people are in favour of this study since they think of it as a step towards a global transparency of educational policies (Meyer & Benavot, 2013). On the other hand, the opponents of PISA raise doubts about the study's validity and reliability (Bolivar, 2011; Duru-Bellat, 2011). Despite these critiques, the PISA study still has an important role to play in the promotion of global educational governance and in the replacement of state sovereignty over educational matters under the influence of international organizations (Meyer & Benavot, 2013).

In 2012, Cyprus took part in the PISA study for the first time in its history and also took part in a new series of research in 2015. The review of the literature has shown that no prior research has been conducted on whether PISA has affected the Cypriot educational system. Therefore, the aim of this research is to uncover the reasons behind Cyprus' participation in the PISA study, its potential influence on the educational system of this country and the ways the PISA study has been exploited by the education policy makers.

This study uses interviews as its research method. The interview participants were selected via criterion sampling in conjunction with snowball sampling or chain sampling (Cohen et al, 2007). In total, sixteen semi-structured interviews of four groups of people were conducted. The four groups consisted of former and current Ministers of Education and Culture, trade unionists, academics and technocrats of the Cyprus Ministry of Education and Culture. The interview data were analyzed based on the method of continuous comparison (Maykut & Morehouse, 1994). In order to validate the data of this research the technique of data triangulation or triangulation 'in the method' was used (Cohen et al, 2008, p. 193) since data were collected from different sources. What is more, a summary of the transcribed interviews was sent to the participants so as to check it (members checks) (Creswell & Plano Clark, 2007).

In summary, the findings of this research show that an important motive for Cyprus' participation in the PISA study was the fact that its participation highlighted and enhanced Cyprus' entity as a state. This is so because Cyprus' participation in this competition was prevented by the veto exercised by Turkey as a member country of OECD. In addition, it seems that the PISA study has been used as a tool for legalizing reforms by the education policy makers (Grek, 2009).

At the same time, it does not seem that the PISA study has affected the Cypriot educational system since the changes that have been made are not the result of this study. Such changes are introduced because the system recognizes the points and practices that need to be changed. As a result, changes conducted on the educational system seem to be adapted to the local context and do not blindly and faithfully follow the OECD policies (Beech, 2009˙ Anderson-Levitt, 2012).

39. The process of social construction of assessment policy: the case of SIMCE

María Teresa Flórez Petour, Jenny Assael Budnik and Cristian Cabalín Quijada (University of Chile, Chile)

This paper draws on a study that promotes a complex and interdisciplinary approach to the understanding of policy processes. On the basis of the three-dimensional approach proposed by Bowe, Ball, & Gold (1992) in relation to the process of social construction of policy, the paper addresses the national curriculum assessment system in Chile (SIMCE) from a holistic and critical view. The trajectories of policy around SIMCE are reconstructed considering: 1) the context of influence; 2) the context of production of policy texts; and 3) the context of practice.

The aims of the study were:

General aim
To understand the process of social construction of policy around SIMCE, from a systemic, historical and interdisciplinary approach, that takes the complexities of the new governance context into account.

Specific aims

- To understand and describe the context of influence of policy around SIMCE, identifying actors, discourses and interactions in the definition of policy.
- To understand and describe the context of production of policy texts around SIMCE, identifying the actors, discourses and interactions in the definition of policy.
- To understand and describe the context of practice in its process of interpretation and translation of policy around SIMCE in school communities, analysing the impact of policy in practice.
- To reconstruct and interpret the process of production of policy around SIMCE, considering the networks of agents involved in the whole process, as well as their interactions and the discourses that circulate between them.

The context of influence, that is, the process of mediatisation of education policy by interest groups, politicians, supranational organisations and other agents in the field of education, is studied by means of analysing discussions around SIMCE in two major newspapers throughout 2014. The context of production of policy texts is studied by analysing policy documents (laws, regulations, decrees, 'expert' and policy brochures, among others) from the perspective of Critical Discourse Analysis and through interviews with policy actors who held key roles in different phases of the development of SIMCE. The context of practice is addressed using an ethnographic approach to study the views and experiences around SIMCE in a set of public schools in Santiago. The interdisciplinary team that leads this research aimed at gradually and progressively connecting all this evidence in order to provide a complex model on how education policy is constructed. In this model, the power networks that operate in the policy process as well as the discourses promoted and silenced in this struggle are portrayed as a complex whole. The study contributes to assessment policy research by providing a more overarching landscape for the understanding of the topic, where all relevant voices are considered and where policy processes are shown in their ideological and contested nature.

References

- Bowe, R., Ball, S., & Gold, A. (1992). Reforming education and changing schools. Case studies in policy sociology. London: Routledge.

Session N: Applying assessment data to inform teaching

40. Diagnostic information from national assessment: Exploration of a simple cognitive diagnostic model

Daniël Van Nijlen and Rianne Janssen (KU Leuven, Belgium)

The primary goal of national assessments is to monitor to what extent educational goals are met at the system level. However, policy makers and educational practitioners also want to draw conclusions that are relevant to the process of teaching and learning and make national assessments function as an 'assessment for learning'. Unidimensional Item Response Theory models are well-suited to make general statements about performance (DiBello & Stout, 2007). However, making national assessments more directly relevant for the teaching process may call for different types of model. The goal of the current study is to explore the possibility of applying diagnostic analysis methods on national assessment data.

Data and method

The data stem from a 2008 study of a test on the mastery of the four basic operations of arithmetic at the end of primary education (474 pupils from 41 schools; 46 items). Data were analyzed using the Bayesian Inference for Binomial Proportion model (Kim, 2011), a conjunctive (non-compensatory) model for cognitive diagnostic assessment. A measure for goodness of recovery of the data (GoR) by the mastery pattern on the attributes was added.
A Q-matrix was constructed using the following attributes: 1) addition, 2) subtraction, 3) multiplication, 4) division, 5) translation of verbal context to mathematical context, 6) use of decimal numbers, and 7) long division with remainder (or decimal number as result).

Results

The overall GoR was 78.8%. As a benchmark, we compared this to the expected recovery if all pupils would be awarded a 1-score for all items (given the overall high score on the test).
The recovery shows a small increase above this benchmark (recovery of 75.9%). The GoR for pupils that master all attributes was 82.5% (n=344). For these pupils the model-predicted item response is always a correct response, so all items that were not answered correctly result in a negative slip. GoR for the remaining group of pupils (non-perfect mastery) is 69.0% (n=130). In comparison to the benchmark for these pupils (58.4%) it is a considerable improvement.
For each attribute, proportion mastery (i.e., easiness) and discriminative power, based on the (point-)biserial correlation of the attribute and the total test score (Tatsuoka, Corter & Tatsuoka, 2004) are calculated (Table 1).

Table 1 Attribute discriminative power and mastery

Attribute	N items	Discriminative power	Mastery non-perfect	Mastery group
Addition	9	.38	.96	.86
Subtraction	8	.51	.93	.74
Multiplication	12	.50	.92	.72
Division	17	.62	.89	.60
Translation context	18	.57	.94	.78
Decimal numbers	23	.67	.87	.54
Remainder	7	.63	.84	.42

Pupils with the non-perfect mastery pattern do not experience much more difficulty with addition. Performance on the division items, however, is considerably lower and certainly the items with a remainder or a decimal number as a result pose problems. These results are in line with the discriminative power which is moderate for addition but fairly high for division.

Conclusion

These analyses can be a valuable addition to the reporting on national assessments. They can yield valuable information for the school-level feedback and introduce some kind of formative element into national assessments.

References

- DiBello, L. V., & Stout, W. (2007). Guest editor's introduction and overview: IRT-based cognitive diagnostic models and related methods. Journal of Educational Measurement, 44, 285-291.
- Kim, H. S. (2011). Diagnosing examinees' attributes-mastery using the Bayesian inference for binomial proportion: A new method for cognitive diagnostic assessment (Doctoral dissertation). Retrieved from https://smartech.gatech.edu/jspui/handle/1853/41144?mode=full
- Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of twenty countries. American Educational Research Journal, 41, 901-926.

41. Programme for International Student Assessment (PISA): Results and cha(lle)nges for the Cyprus educational system

Salome Hadjineophytou (Saint Louis University, Cyprus)

The Programme for International Student Assessment (PISA) is an international survey commissioned by the Organization for Economic Cooperation and Development (OECD).
The survey takes place every three years and tests skills and knowledge of 15-year-old students in the key subjects of reading, mathematics and science. Since 2012, students' skills in problem solving are also being assessed. The tests are not related directly to school curricula, but assess students' ability to transfer learning to everyday situations. In addition, the survey collects information on students' family (e.g. parent educational attainment and occupation), and the motivational influences in their home (e.g. computer, internet, educational software, books, dictionaries).

Cyprus participated in PISA in 2012 and 2015. The results of PISA 2015 will be announced in the end of 2015. According to the results of PISA 2012, Cyprus ranked 44th in reading, 46th in mathematics and 50th in science out of 65 countries. The Ministry of Education and Culture of Cyprus reviewed the results and initiated discussions among the main stakeholders, i.e. the Pedagogical Institute, the Centre for Educational Research and Evaluation, and the Directorates of Secondary Education, giving emphasis on the challenges the Cyprus education system has to meet and the changes needed to be made (UNESCO, 2015). The discussions gave emphasis mainly on the early identification of low and high achievers and the way schools should address such students, syllabuses' reviews in relation to cognitive and metacognitive abilities, new teaching and learning approaches on developing problem-solving skills, monitoring and evaluating students' results, structural changes, school culture and finally establishing positive attitudes towards learning process (UNESCO). The above changes and challenges aim for students to perform better in international surveys such as PISA, especially when the surveys focus on assessing the transfer of learning in everyday situations.

In relation to parent educational attainment and occupation, as well as the motivational influences in students' homes, the results revealed that students belonging to higher socio-economic status showed better performance than students from lower socio-economic status (OECD, 2014). Additionally, those parents who had higher expectations, motivated and guided their children so that they showed better performance. A percentage of students from low socio-economic status showed high performance and these students are defined as resilient; however, the percentage of Cyprus was very low (1.9%), whilst in Korea, Shanghai and Hong Kong was more than 13% (OECD). Due to the financial crisis Cyprus faces the last few years, the education system has to meet another challenge, which is to provide students from low socio-economic status motivational influences that their parents cannot. In addition, the Cyprus education system needs to increase the percentage of resilient students, as according to OECD (2011), schools may have a significant role to play in order to promote resilience, by increasing time in class, motivation and self-confidence.

In conclusion, the results from PISA 2012 revealed a number of challenges for the Cyprus education system. In order to proceed to the necessary changes, the system needs to focus on various areas that will enable students to transfer their learning to their everyday situations. Finally, the system needs to focus even more to those students that are less privileged.

References

- OECD (2011). How do some students overcome their socio-economic background?
- PISA in Focus, 5. OECD Publishing, Paris.
- OECD (2014). PISA 2012 Results in Focus: What 15-year-olds know and what they can do with what they know. OECD Publishing, Paris.
- UNESCO (2015). Education for All 2015 National Review Report: Cyprus.

42. MathemaTIC – a multilingual, digital and adaptive environment for increasing equity in mathematics learning in Grades 5 and 6 students in Luxembourg

Philippe Arzoumanian(DEPP, Ministry of Education, France) and Amina Kafaï-Afif (Agency for Development of Quality in Schools, Ministry of Education, Luxembourg)

With almost 45% of foreigners among its residents, Luxembourg occupies the top position among the EU states with respect to the proportion of students who do not speak the language of instruction at home, the number of foreign languages taught at school and the number of hours dedicated to the teaching of these foreign languages. This multilingual tradition heavily impacts national student performance and equity, evidence which is supported (in PISA studies since 2000 and national standardized tests since 2008) by the performance gaps observed between students from various socio-economic backgrounds, between those of native and foreign origin as well as between boys and girls. In response to its growing heterogeneous school population, the Ministry of National Education, Children and Youth (MENJE) introduced MathemaTIC in September 2015, a project whereby 10-11 year-old students in grades 5 and 6 of 40 primary schools are able to learn mathematics in four languages in an adaptive digital environment. The fundamental aim is to increase the chances of these students succeeding in Mathematics, using a technology-enriched solution while eliminating language as a barrier to learning.

Since the past few decades, the school population of Luxembourg has experienced an increased heterogeneity in terms of migration, language and social origins. Offering schools a 'language-free' solution to better support classroom learning is one of the numerous national strategies to tackle inequity, arising among the growing proportion of students who struggle to master the language of instruction which is indispensable for them to succeed in their learning. MathemaTIC offers a digital adaptive learning environment with items in German, French, Portuguese and English. It is based on the national curriculum which enables personalization of the student's learning path and instruction in mathematics. The underlying technology enabling real-time feedback during learning, consequently modifies the mode and purposes of assessment and opens up the opportunities for students to succeed in school.

These advantages however come with a fair share of methodological, technical and practical challenges linked to the assessment itself and the corresponding data generated. These include the need to create reliable mathematic items which are pedagogically suitable and as well in line with didactical research, political orientations and aligned to the curriculum. The various sources of formative assessment data need to be matched to create automatic student profiles and learning paths. The big data which result need to be filtered, reduced and structured to offer teachers meaningful information that subsequently feed into the instruction. This form of personalized and differentiated teaching and learning thus alters the largely prevailing traditional classroom model. Professional teacher development programs hence need to be organized to support teachers to gradually integrate this innovative approach in their everyday practice.

Introducing this adaptive learning environment does not only relate to a measurement issue, but it affects the equity of teaching and learning. The multilingual aspect of MathemaTIC gives a fairer chance to students who do not understand the language of instruction and whose achievement would therefore otherwise not be measured well when the language becomes a barrier to understanding the mathematical text. This disadvantage is extended when the mathematics results are used to make a variety of decisions, from promotion to inclusion in special programs. As MathemaTIC is aimed at adjusting to the ability of individual learners, it reflects their current achievement level, hence empowering both the students who are falling behind and those who sail through.

In our presentation, we will illustrate how MathemaTIC contributes to raising equity by reworking the traditional teaching, learning and assessment paradigm so as to reduce the performance gap between high and low achievers in mathematics.

Session O: Assessing what matters – validity challenges in assessment

43. A new approach to the reform of vocational qualifications in Wales by Europe's newest qualification regulator

Cassy Taylor (Qualifications Wales, United Kingdom)

Qualifications Wales was established in autumn 2015. It has two principal aims:

- Ensuring that qualifications, and the Welsh qualification system, are effective for meeting the reasonable needs of learners in Wales; and
- Promoting public confidence in qualifications and in the Welsh qualification system.

Upon the point of Qualifications Wales' establishment a programme of reform of general qualifications was underway, inherited from Welsh Government. Less progress had been made in relation to the reform of vocational qualifications but Qualifications Wales was keen to redress this balance and appointed Cassy Taylor as Associate Director (Vocational Qualifications) to oversee this work.

Rather than develop a 'one size fits all' approach to vocational qualifications, Qualifications

Wales instigated a long-term programme of sectoral reviews. The ‘reasonable needs of learners’ differ between vocational areas. Without getting ‘under the skin’ of each sector key issues, that need attention to make qualifications more effective, may be missed.
Qualifications Wales began with the Health and Social Care sector– the largest employment sector in Wales with qualifications that are required for employment.

A model for sector review of analysis – solutions – implementation – impact was developed, with six months allocated to the analysis phase. The objective of this phase was to identify how effective the present qualifications and system are in meeting the needs of leaners and the steps that Qualifications Wales could take towards making these more effective. Through a combination of intensive engagement with stakeholders and the interrogation of data the review team developed an understanding of the qualifications landscape for Health and Social Care and identified where there were weaknesses in meeting the needs of learners.
Matters considered included the sufficiency, range, currency and nature of qualifications; the effectiveness of assessment; the availability of Welsh medium assessment; whether the requirements of employers and higher education are met; whether the qualifications are comparable with those elsewhere; and efficiency and value for money.

The review team undertook over 125 engagements with stakeholders, including schools, further and higher education and work based learning, as well as employers, sector bodies, government agencies, service users and commissioned focused discussion groups with over 800 learners.

The review has identified a number of questions relating both to the qualifications and the qualification system in the sector. Stasz (Stasz 2011) noted that 'there is a moral hazard problem for providers who also act as assessors where the NVQ is the unit of funding and where funds are dependent on its achievement' and the Review draws attention to the complexity of roles of different agencies in the system and raises similar questions about whether the existing learning and assessment delivery system across further education and work based learning is optimised to support effective assessment.

Questions are also raised about the effectiveness of the design of qualifications designed to assess competence in the sector and observes that, to a point, the occupational standards have been used as a substitute for robust and targeted assessment design. In combination, the issues identified have resulted in a determination to engineer, through the appropriate use of regulatory powers and through forging new relationships with players in the system, a new, more holistic and innovative approaches to competence assessment which consider the extent to which sufficient confidence can be placed in competence rather than on a 'total coverage' philosophy.

The findings of the review have been positively received by the sector and the potential improvements to qualifications and the system which will follow seem likely to be far-reaching and significant.

References

- Stasz, Cathy (2011). The purposes and validity of vocational qualifications. Skope Research Paper No. 105.

44. How valid are pre-university vocational qualifications?

Rose Clesham (Pearson, United Kingdom)

Students in schools and colleges in England can follow two routes to gain admission to higher education or university. For general academic qualifications in the UK (GCE),content standards and assessment objectives are provided to awarding bodies by the government and regulator to provide consistency in terms of required cognitive knowledge necessary in subjects to attain A level qualifications (General Certificate of Education).

Alongside general academic qualifications, there are a range of pre-university Level 3 vocational qualifications available that have dual purposes. Students studying these courses can use them to gain admission to university degree courses or they can use them to move directly into the workplace or apprenticeship schemes.

There are two particular aspects of L3 vocational qualifications that make them significantly different from general academic L3 qualifications:

- they are work related qualifications and therefore students learn and are assessed on a range of knowledge and skills appropriate to a particular vocational sector;
- At present, in the Qualifications and Credit Framework (QCF) qualifications, assessment is entirely based on an internal assessment model, using verification procedures to quality assure and maintain the standard.

This research study investigated a range of UK pre-university QCF L3 vocational qualifications in terms of the knowledge and skills they assessed, at what levels of demand and how aligned the assessments were in relation to the intended curricula and instruction for such courses.

Validity is the central concept underpinning the development and evaluation of qualifications and their associated assessments. Two of the key components of educational validity, and of particular focus in this research study are that there is alignment between the intended curriculum and the assessments and that the content standards and assessments have the appropriate breadth and depth in terms of the level of the qualification and its intended purpose (Kane, 2009). There are various forms of curricula, intended (usually content standards), enacted (what actually happens in the classroom), assessed (the tests/examinations) and learned (ultimately what students know and can do). This research was predominantly an empirical study, involving experts in vocational education mapping the assessments of pre-university vocational qualifications from a range of sectors, to consider the type of knowledge and skills being assessed and their relative emphases.

A key research question was how aligned were the intended and assessed curriculum- what type of knowledge did the content standards of vocational qualifications expect students to experience and learn, and how well did the assessments align and reflect these traits?

Due to the applied nature of vocational qualifications, and their assessment methodologies (mainly using criteria referencing), their assessments can offer the potential to assess cognitive (knowledge based), affective (attitudinal/behavioural) and psychomotor (can-do, competence) skills. These skills, and the differing methods by which they are evidenced, are the basis for all inferences, classifications and decisions taken about the knowledge and competencies of individuals or institutions (Madaus, 1992; Griffin, 2007).

The analyses of these qualifications required a methodology to map all of these aspects, and provide data profiles of what was being assessed, and at what level of demand. This methodology will be described and research outputs and findings shown. Finally, the implications concerning the design and development of vocational qualifications will be discussed.

References

- Griffin,P. (2007). The comfort of competence and the uncertainty of assessment. Studies in Educational Evaluation, 33, 87-99.
- Kane. M. (2009) Validating the Interpretations and uses of test scores in Lissitz (Ed) The Concept of Validity: Revisions, New Directions, and Applications.p39-64.
- Madaus, G.F. (1992). An independent auditing mechanism for testing. Educational Measurement: Issues and Practice, 11(1), 26-31.

45. Digital responsibility – a required literacy for citizenship: How to understand and measure the concept?

Ove Edvard Hatlevik and Inger Throndsen (University of Oslo, Norway)

The emergence of Information and Communications Technology (ICT) has influenced how people organise and carry out working life and social activities. ICT are also influencing the educational systems in Europe, and there are different endeavours to implement ICT in schools. As students get more frequent access to ICT this probably will change how they search for information, how they share information and communicate with others. It is therefore important that students are able to make safe and responsible judgements about how to use ICT in their learning.

Since 2006 the ability to use ICT, i.e. students' ICT literacy, has been one of five competence areas in the Norwegian curriculum. In the Norwegian framework for basic skills, ICT literacy is operationalized into four sub-categories: communication, production, searching/processing information, and digital responsibility.

Digital responsibility deals with students' capability to make safe and responsible judgements when they are involved in ICT use, e.g. knowing about rules for privacy. Digital responsibility is a required literacy as students are facing challenges related to privacy protection and information security through their use of online resources and communication with other people. As digital responsibility is a sub-category in the national curriculum, it is expected that all students will be given equal opportunities to develop these skills and literacies.

In 2013 the International Computer and Information Literacy study (ICILS) (Fraillon et al., 2014) was conducted in 21 countries, including Norway. The students completed a proficiency test measuring different aspects of ICT literacy, including safe and secure information use. Some argue that students acquire ICT literacy by themselves. However, assessments of ICT literacy reveal differences between students. Therefore it is highly relevant to examine how digital responsibility is understood and measured.

This paper addresses the following research questions:

- How are the items measuring using information safely and securely in ICILS 2013 related to the content of digital responsibility in the Norwegian curriculum?
- Based on data from ICILS 2013, what kind of items measuring using information safely and securely are Norwegian students able to solve, and what kind of items are challenging?

This paper has three sections. First, the descriptions of digital responsibility in the Norwegian curriculum and using information safely and securely in the ICILS framework will be analysed. Second, items from the ICILS test will be mapped to the description of digital responsibility in the Norwegian curriculum. Third, we use the test results to identify items that students have answered in a correct or wrong way.

It seems that the concept digital responsibility in the national curriculum has many similarities with how using information safely and securely is described in the ICILS framework. Both concepts are dealing with making judgements about using ICT. When it comes to measuring digital responsibility in ICILS 2013, we identified ten tasks related to this concept. These tasks are for example dealing with protecting identity, protecting information and identifying what is suspicious in emails. The tasks are in line with descriptions of digital responsibility in the curriculum, but there are also themes in the national curriculum that is not covered by the test. Preliminary findings show that students are rather successful with tasks focusing on protecting explicit personal information. This might be due to the emphasize schools, parents and peers place on this kind of knowledge. However, the results also indicate that students struggle with items about suspicious emails, and one reason could be that students are using other ways of communicating (i.e. messages and apps).

References

- Fraillon, Ainley, Schulz, W., Friedman, T., & Gebhardt, E. (2014). Preparing for life in a digital age. Amsterdam: IEA.

Session P: Assessment and learning in the digital age

46. Measuring the Impact on Learners: evaluating a whole school reading programme in the UK

Grace Grima, Elpida Ahtariou, Krystina Dunn, Vanessa Greene (Pearson UK, United Kingdom), Sue Bodman, Glen Franklin, Jane Hurry and Catherine Carroll (UCL, United Kingdom)

The pedagogy of this whole school reading programme is built on the seven-year Clackmannanshire study, the findings of which show that systematic synthetic phonics is an effective way to teach children to read. This resource, which is used in more than 5,000 primary schools in UK is designed to engage a generation of children used to reading/playing online, and brings together 350+ books for different reading levels, quizzes and games, in print and online.

The evaluation of the resource comprises a randomised control trial (RCT) and a process evaluation, it spans over five scholastic terms (Jan 2015-Jun 2016). This multi-strategy study was designed to explore the impact and implementation of the resource, in particular:

- The impact on pupils' literacy learning, their attitudes to reading, school and their reading activity.
- The wider impact on pupil, teacher, school and parent outcomes.

This is one of a few RCTs of a whole school reading programme undertaken with pupils in Years 1 and 2. (ages 6-8 years) Complementary data collection included interviews and observation data from 10 case study schools which explored in more detail the findings from the surveys and pupil assessment data.

Schools were randomly allocated to one of two groups:

- Intervention group: 21 schools which had the resource in January 2015–July 2016.
- Control group: 15 schools which had the resource from January – July 2016.

The sample of 1,510 pupils on which these interim findings are based, were drawn from 30 matched schools.
The University of Durham InCAS standardized assessments were collected at baseline (January 2015) and at first follow-up (June 2015) from all intervention and control schools.
In addition, a pupil self-report to measure the impact of the resource on pupils' literacy learning, attitudes to reading and school and their reading activity and data relating to implementation and usage of literacy materials and attitudes to teaching and learning were gathered. Teacher surveys of usage were collected monthly. These data were collected in 30 matched schools randomly allocated to either an intervention or control group

All schools were involved in the process evaluation. Key evaluation areas included impact on pupils, schools, teachers and parents in the intervention schools, pupil, parent and teacher perceptions, and patterns of usage and implementation. Questionnaires relating to attitudes to literacy learning and levels of confidence were collected for all Year 1 and Year 2 teachers in both intervention and control schools. The experiences of 10 case study schools were investigated in more depth as they implemented the reading resource.

The interim findings of the research will be shared during this session and the discussion will focus on challenges and opportunities offered by such research that link curriculum, teaching, learning and assessment.

References

- Bodman, S. & Franklin, G. (2014) Which Book and Why: Using Book Bands and book levels for guided reading in Key Stage 1. London, UK: UCL Institute of Education Press.
- Ciampa, K. (2012) The effects of an online reading program on grade 1 students' engagement and comprehension strategy use. Journal of Research on Technology in Education, 45(1), 27-59.
- Hurry, J. & Sylva, K. (2007). Long-term outcomes of early reading intervention. Journal of Research in Reading, 30, 227-248.
- Hurry, J., Sylva, K. & Riley, J. (1999) Evaluation of a focused literacy teaching programme in reception and Year 1 classes: Children outcomes. British Educational Research Journal, 25, 637-649.
- National Literacy Trust (2013, May 16). Children's on-screen reading overtakes reading in print. Retrieved from http://www.literacytrust.org.uk/news/5372_children_s_on-screen_reading_overtakes_reading_in_print

47. Exploring the effects of undertaking the Extended Project Qualification

Charlotte Stephenson (AQA, United Kingdom)

Externally marked examinations are the default method of assessment in England, with current educational policy restricting the amount of school-based assessment within each general qualification (Department for Education (DfE), 2015). The exact evidence-base for this shift in policy is unclear, but has been attributed to the burden school-based assessments place on teachers, their inherent bias, and their susceptibility to malpractice (DfE, 2012). The movement away from school-based assessment has occurred despite wide acknowledgement that it enables the measurement of skills missed by examinations, including practical and oral skills, creativity and reflective thinking (Secondary Examinations Council, 1986). Furthermore, internal assessment has been found to increase students' motivation and sense of responsibility whilst stimulating discussion and imagination (QCA, 2005).

The term 'school-based assessment' encompasses a range of assessment types, including those that are project-based, such as the Extended Project Qualification (EPQ). The EPQ is one of few remaining academic qualifications in England comprised entirely of school-based assessment. It was introduced during a UK government curriculum reform effort to combat 'disengagement and underachievement' (Working Group on 14–19 Reform, 2004, p. 4). The qualification was intended to enable learners to develop and demonstrate research and analysis skills, problem solving, critical thinking and presentation skills through project-based learning (PBL) (Working Group on 14-19 Reform, 2004).

PBL is a curriculum model in which students learn through undertaking complex tasks, often based on solving real-life problems or answering challenging questions (Thomas, 2000). PBL is believed to promote learning by connecting new knowledge with prior experience, increasing engagement, self-direction and motivation (Filippatou, 2010). Through PBL, skills such as problem-solving ability, critical-thinking skills and organisation skills have been found to be enhanced, and academic performance improved (Blumenfeld et al., 1991; Boaler, 1998; Thomas, 2000); these are just some of the skills the EPQ was intended to promote (Working Group on 14-19 Reform, 2004).

In a recent paper, Jones (2015) found that students taking the EPQ with AQA showed enhanced performance in English, Business, Sciences, Art and Humanities A levels after controlling for prior attainment. This paper reports on a qualitative research project that explores the possible underlying mechanisms that contribute to this effect, and, in doing so, highlights the potential benefits of PBL.

Whilst literature on PBL and its effectiveness exists, the majority explores its effects on academic performance within the same academic discipline as the project undertaken, e.g., the effects of undertaking mathematics projects on performance in mathematics (Boaler, 1998). Studies that have explored the effects of PBL on general academic performance tend to be in the context of schools that have undertaken a whole-school project-based reform effort (Thomas, 2000). This paper contributes to the extant literature, in partnership with Jones' (2015) aforementioned quantitative research, by exploring teachers' perceptions of the effects of undertaking a project-based qualification, specifically the EPQ, on students' academic performance in other subjects and their academic performance more generally.

To explore their experiences, focus groups were held with 30 teachers: five focus groups with six teachers in each. Semi-structured interview schedules were used with open questions to elicit detailed responses. The data were analysed using thematic analysis to identify common themes and member checking was used to ensure results were aligned with teachers' experience. The findings elucidate the possible underlying mechanisms that contribute to the effects of the EPQ observed by Jones (2015), and in doing so, highlight the potential benefits of PBL and inform how this link could be investigated quantitatively. Further, this research captures teachers' views of school-based assessment at a time when examinations are the preferred approach to educational assessment in England and school-based assessment has been limited (DfE, 2015).

48. Investigating student-student interactions in an assessment of collaborative problem solving: An in-depth analysis of think-aloud protocols

Ronny Scherer and Fazilat Siddiq (University of Oslo, Norway)

Computer technology has evolved considerably and has become pervasive in our society. Moreover, students take the ubiquitous access to the Internet and digital tools for granted and tend to be 'always on' (Oblinger, 2004). These changes have affected the educational practices and are consequently forcing the assessment systems to follow. Several agencies have stressed the importance of 21st century skills as critical for functioning in education and work life.

The notion of 21st century skills refers to a set of skills such as communication, collaboration, critical thinking, information literacy, problem solving, and others (Binkley et al., 2012); particularly students' collaborative problem solving (ColPS) skills have been highlighted as critical. Nevertheless, only a limited number of studies have investigated students' ColPS despite its importance and relevance (Siddiq, Hatlevik, Olsen, Throndsen, & Scherer, 2016). This study aims to address this gap by investigating the processes involved in student-student interaction while solving tasks within a computer-based test environment. In particular, students' competences in solving problems collaboratively and communicating with peers in digital environments are explored.

The ColPS task was developed as part of an assessment of ICT literacy. It was constructed as a task that does not require specific content knowledge; in fact, this feature increased the possibility for all students in a team to contribute to the collaborative part. Moreover, students were grouped in teams of four and took the test simultaneously but were physically apart. They first read a poem and watched a YouTube video related to the poem. In the ColPS task, the students were then asked to join their team and make a sketch/drawing which should express their interpretation of the poem together with their team members.

Students' screen activities and think-aloud protocols were recorded. A coding scheme was developed and the data were coded and analyzed focusing on student-student interactions, role-taking, and problem solving processes. The two groups comprised four ninth-grade students. Each student was interviewed after the test to obtain more information about the use of and experience with digital communication and collaboration in classroom activities.

Our results indicate that students' ColPS skills are closely related to their experience with ICT and that they to a large degree struggle with digital communication and collaboration. Furthermore, we found varying patterns of collaboration among group members in a group that were related to students' role-taking and participation. These patterns will be discussed in light of the specific needs for instruction and learning that promote students' ColPS skills. We finally point to consequences for designing and evaluating appropriate ColPS assessments, focusing on student-student interactions. We believe that it is vital to monitor them for providing students equal opportunities to develop ColPS skills.

References

- Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining Twenty-First Century Skills. In P. Griffin, B. McGaw, & E. Care (Eds.), Assessment and Teaching of 21st Century Skills (pp. 17-66): Springer Netherlands.
- Oblinger, D.G. (2004). The Next Generation of Educational Engagement. Journal of Interactive Media in Education, 2004 (8). DOI: http://doi.org/10.5334/2004-8-oblinger
- Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I., & Scherer, R. (2016, accepted). Taking a future perspective by learning from the past–A systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. Educational Research Review.

Session Q: Good and bad consequences of assessment

49. Teacher evaluation – trapped between accountability and learning: Assessing teacher professionalism – formatively

Sølvi Lillejord, Kristin Børte (Knowledge Centre for Education, Norway) and Therese Hopfenbeck (Oxford University Centre for Educational Assessment, United Kingdom)

In this paper we ask if formative teacher evaluation may contribute to strengthening teachers' understanding of assessment for learning. Behind this question is the observed paradox that while teachers are expected to use assessment formatively in their teaching, the same requirement does not apply when districts and school leaders evaluate teachers' work.
A systematic review (Lillejord et al. 2014) shows that many systems for teacher evaluation aim simultaneously at control (quality assurance) and improvement. Research shows, however, that systems fail to achieve this double goal and that many teacher evaluation practices instead are reduced to inadequate summative accountability (Marchant et al (2015). The review reveals that systems fail when they are driven by an administrative logic and ignore the following central insights from assessment research: 1) provide enough time and resources to conduct assessments, 2) make sure the assessment is valid and credible, 3) clearly define the object of assessment, 4) linking assessment to improvement and 5) minimize unintended effects (Assessment Reform Group 2006).

Why this happens, and how, will be exemplified by research on teacher evaluation systems and practices in Chile, China and Portugal. The paper highlights four generic challenges in teacher evaluation systems related to violation of these principles.

Firstly, when systems for teacher evaluation require extensive documentation, insufficient time and resources is set aside for the evaluation. Research shows that both teachers and evaluators struggle with this problem. As a result, the assessment process was not completed or followed up, and was increasingly perceived as a burden by all actors. Secondly, teachers expect assessors to be highly qualified educators, with teaching experience and expertise. When this expectation is not met, teachers don't consider the results of the assessments as valid and reliable because they find it difficult to trust the assessors. Thirdly, evaluation systems were complex because they combined several approaches based on qualitative and quantitative data. As a consequence, the object of assessment became unclear. When the object is unclear, it is also uncertain what has actually been assessed. A fourth challenge was that some of the systems were overly bureaucratic in the sense that they approached teacher evaluation with an administrative knowledge interest and with meagre understanding of feed-back loops and learning processes.

These four challenges may explain why systems and practices investigated do not achieve their double goal of control and improvement. Instead of formatively supporting teachers' professional learning, the assessment ended up being summative. One exit strategy from what seems to be a gridlock situation is to insist that teachers are knowledge producers and schools sites for knowledge development. Increasingly, researchers stress that teachers should use student feedback to adjust their teaching (Darling-Hammond 2016). Such practices should be studied and documented by teachers as well as researchers. The deficit perspectives underpinning many teacher evaluation systems should be replaced with a resource-perspective. Generations of teachers have been assessing their pupils. Formative assessment is integral to instruction and central to teachers' professional competence. Hence, assessment competence resides in schools. However, reviews of assessment for learning practices show that even though teachers give pupils feedback on their learning, only few empirical studies investigate how pupils give feedback to their teachers on how they perceive the teaching. As few schools or countries have documented systems where teachers improve their teaching based upon feedback from their pupils, we need more knowledge about how teachers use input from various sources to adjust or improve their teaching. Providing teachers with systems for how to improve their teaching, by including teachers and pupils in a systemic practice, could be one way of moving teacher evaluation forward.

50. Understanding and Developing Relational Aspects of Assessment for Policy and Practice

Ruth Dann and Jo Basford (Manchester Metropolitan University, United Kingdom)

The aim of this paper focuses on articulating a relational approach to assessment drawn from research focused on assessment practices in the context of assessment for learning. The research methods draw from two different case study projects, which are synthesised in order to develop a new understanding of assessment as a relational process. One case study focuses on the practices and perceptions of (5)teachers in relation to their understanding and enactment of assessment in an early years context. The second, explore the ways in which eleven children, who were underachieving (aged 10), understand feedback and the next steps of their learning.

This paper is contextualised within international debates surrounding the impact of global standardised testing and its effect of distorting teaching and learning relationships and narrowing the curriculum. Furthermore, it is located within more recent policy changes in England, in which ‘Assessment without Levels’ calls for new ways of imagining and enacting assessment practices.

The research findings show that ‘teacher assessment’ practices were distorted as the focus for teachers’ pedagogical interactions became outcome orientated rather than focused on the learning process. Teachers reported their priority to conform to perceived policy interpretations of effective assessment. When exploring the feedback process with pupils, findings revealed that the children had not internalised the feedback given. Few pupils (aged 10) could identify any next steps for their own learning in mathematics. In literacy, they were able only to mention low level next steps for learning. It was clear that teachers were assuming that these pupils would accept the carefully constructed and uniform feedback given as part of whole school policy, in particular ways.

This research is framed within a socio cultural pedagogical perspective. The research findings are examined theoretically through the use of Bourdieu’s notions of ‘field’ and ‘habitus’ which offer insights into individuals’ interpretations of their positioning and the cultural and social capital they use to operate within their field (Bourdieu, 1998). Particular attention is paid to the power relations of different ‘players’ in the educational ‘field’ (pupils, teachers and policy makers) and the consequences of such relationships for assessment practice. Ways in which this is ‘played out’ are closely examined in order to reveal the nature of these conforming relationships. This paper is particularly located in a space which identifies the impact of policy at national and school levels and teases out their impact on specific professional and pedagogical relationships. This is well aligned to the theme of this conference.

The evidence from both case studies concludes that 'pseudo' relationships form an extensive part of the assessment practices examined. These are built on conforming positions in order to fit in with perceived policy expectations (Mac Naughton 2003). Beyond such relationships, this paper argues for and develops a transformative notion of assessment as a relational process both in terms of enactment at the interface of teaching with pupils and between teachers. The development of assessment as a relational process, draws pedagogically from Biesta's (2015) notion of a relational pedagogy, professionally from (Wenger, 1998) conceptualisation of a professional community of practice, and socio-culturally from Vygotsky (1962). This research offers a new way of framing assessment as a relational process. It argues that the relationship between teachers and pupils in the process of assessment for learning should be reconceptualised as a new understanding of the cognitive gap between what pupils know now and need to know next. Rather than this being predominantly teacher pre-determined, it should recognise and provide opportunities for pupils' own considerations and articulations of this gap. This is conceptualised within a socio-cultural pedagogy and 'assessment as learning' (Dann 2002).

51. The teacher as a stakeholder in utilizing an assessment for learning tool

Guri A. Nortvedt and Anubha Rohatgi (University of Oslo, Norway)

Early identification of at-risk students and assessment for learning (AfL) are emphasised in national educational policies in Norway and thus influence educational assessment programmes. One of the tools provided for primary school teachers are mapping tests aimed at identifying students who score under the threshold judged to be necessary for acquiring adequate numeracy competence. The annually administered tests assess students' number concept and calculation skills. The tests have a ceiling effect by design. A student at the 20th percentile typically answers approximately 75% of the items correctly. The underlying rationale is that the mapping tests should be used for AfL and should provide teachers with evidence of what identified students can typically accomplish and serve as the basis for planning teaching interventions.

Even though AfL has been the gold standard for some time, the uptake of AfL practices has proved challenging for teachers (Black & Wiliam, 2012). Previous research indicates that principals' involvement, shared accountability and opportunities to cooperate with peers might help teachers develop a positive attitude towards to the use of assessment and thus support student learning (Goddard, Goddard, Kim, & Miller, 2015; Hallinger & Heck, 2010).

The responsibility for identifying where students are in their learning processes and what needs to be done to move them forward lies primarily with the teacher (Black & Wiliam, 2012). Educational authorities trust teachers to prepare for and administer the tests as well as score and interpret results for their students. Test results are owned by the local school. They might be shared with local school authorities but are not reported nationally. As such, the test are considered low-stakes.

This paper report on a Norwegian study, comprising semi-structured interviews and online surveys with primary school teachers and principals about how they prepare for the mapping tests, administer and interpret test outcomes and plan interventions. The analysis will focus on the challenges and expectations the teachers face in their work with the mapping test.

Analysis indicates that principals trust their teachers to administer and score the mapping tests according to national guidelines. Teachers share test outcomes with their principal. While shared accountability is observed in some schools where teachers and principals interpret test outcomes together, more hierarchical systems are observed in other schools, perhaps leaving too much responsibility to the teachers. The teachers show an interest in and concern for the outcomes of the assessment and treat the test as a tool that can help them improve their teaching. However, some principals report that teachers in their school administer practice tests items with students prior to administration of the actual tests in non-acceptable ways. Even more alarmingly, the interview data revealed that some teachers find interpreting the results and planning interventions rather challenging and so delay these activities until subsequent semesters. We propose shared accountability and what Heck & Hallinger (2010) label 'collaborative decision making' as a way forward to help teachers utilize the mapping test as an AfL tool as intended.

References

- Black, P., & Wiliam, D. (2012). Developing a theory of formative assessment. In J. Gardner (Ed.), Assessment and learning (2nd ed.) (pp. 206–229). London: Sage.
- Goddard, R., Goddard, Y., Kim, E. S., & Miller, R. (2015). A theoretical and empirical analysis of the role of instructional leadership, teacher collaboration, and collective efficacy beliefs in support of student learning. American Journal of Education, 121(4), 501–530.
- Hallinger, P., & Heck, R. H. (2010). Collaborative leadership and school improvement: Understanding the impact on school capacity and student learning. School Leadership & Management, 30(2), 95–110.

Session R: Social and political context of assessment

52. Social, political and cultural impact of high-stakes national assessments: the case of UNT in Kazakhstan

Aigul Yessengaliyeva (PhD Sociology, Kazakhstan) and Nico Dieteren (Cito, The Netherlands)

In modern Kazakhstan the main high-stakes assessment is UNT (Unified National Testing). UNT combines final certification of graduates of institutions of secondary education and entrance examinations to higher professional education. UNT is carried out for graduates who wish for the current school year to enter higher education of Kazakhstan. This type of assessment was implemented to the national educational system twelve years ago and it came as an alternative for the old Soviet assessment system.

Implementation of a new type of high-stakes assessment had significant social and cultural consequences for the society of this new independent country. During these years there was formed a new generation of Kazakhstani with different views, ways of thinking and attitudes behavior. These changes had a contradictory character. There are many positive impacts for social values and views. But on the other side the new high-stakes assessment has given rise to new problems and led to some negative consequences.

Based on published materials and own survey this paper will present an analysis of the social, cultural and political influence of high-stakes assessment on the Kazakhstan society.
This influence will be considered through the past, present and future perspectives. The own survey included narrative interviews with different stakeholders of this assessment process: teachers, parents, students, professors of universities and employers.

Introduction of the UNT has brought advantages to Kazakh education and society: less subjectivity and more transparency in assessment; more opportunities to compare and evaluate schools and teachers; more social justice and more equal chances for candidates from non-urban regions and from lower social classes.

There are still elements to improve. High social pressure brings about negative social reactions like corruption, dishonesty and more suicide attempts. The quality of the assessment can be improved, as most questions are quite factual and reproductive. Alignment of the UNT with skills and attitudes needed for successful university career still has to be improved. (ref: Winter, L e.a. (2014), The culture and practice of assessment in Kazakhstan, in: Educational Reform and Internationalisation: the case of school reform in Kazakhstan, ed. Bridges D.)

Currently this situation of positive and negative outcomes of the UNT is one of the main policy targets of the Kazakhstani government, when it comes to modernize Kazakh education and make it ready for the 21st Century.

In the Netherlands, high stakes final examinations that combine final certification of graduates for secondary education and qualification examinations for entering higher education, have a longstanding history of almost 45 years. As in Kazakhstan, these final exams have a very high social and political impact. Negative effects like high stress, training to the test and discussions about objectivity and transparency are common in Dutch society. Up to the late nineties in the previous century, these final exams were predominantly assessing reproductive knowledge and very limited some skills. Institutes of higher education were more and more complaining about a bad connection between the qualifications assessed in central exams and the skills and study attitude needed to be successful in higher education. Starting in 1999 central exams and the curriculum were changed: more focus on skills and less focus on knowledge, called 'Studiehuis'. Recently Dutch government has commissioned a group of experts from higher and secondary education, together with representatives of the enterprises, to build a new vision on secondary education and how to prepare students for studying and working in 21st Century: 'OnsOnderwijs2032' (ref: see www.onsonderwijs2032.nl). We will discuss consequences for the high stakes central exams in the Netherlands.

Further research will be proposed to see how modernization of UNT in Kazakhstan can learn from previous experiences in the Netherlands.

53. Four Issues in the Debate about Admissions Testing and the Four P's: Psychometrics, Politics, the Press and the Public

Yoav Cohen and Anat Ben-Simon (NITE, Israel)

Educational assessment does not operate in a free environment. It is tightly connected to the well-being of the members of a society and cannot be divorced from social considerations. The establishment of high-stakes admission tests to higher education on a national level, or a reform in an existing system (e.g. Balf 2014), invariably starts an ongoing public and political debate. It involves the academe, the public, the legislature and the testing community in a debate about the costs and benefits to the institutions and the social groups that are involved. In this paper we present the main issues of this debate, which involves politicians and the academic establishment, as well as the public and the press. Other players in the arena of admissions testing are various NGO's that take it upon themselves to serve and represent disadvantaged groups. They communicate their concerns directly to the academe, and indirectly through politicians and the press.

The main issues in the public debate are the effect of coaching on the test, the predictive validity of the test, the fairness of the test in relation to various social groupings (according to gender, socio-economic status and ethnicity), as well as questions relating to the contents of the test.

The interplay between the academic world and the political establishment, which is monitored closely by the press, reveals quite a few misunderstandings and differences of opinion (on all sides) regarding the role of the academe, admissions testing, the science and technology of psychometrics and the definition and meaning of fairness. In spite of the extensive research programs carried out and published by psychometricians and educational researchers, both politicians and journalists keep iterating erroneous factoids about testing in general and admissions testing in particular. Misinformation about admissions testing is also shared by members of the academia. Thus, for example, despite numerous research reports that reveal good predictive validity of admissions tests (e.g. Kennet-Cohen, Bronner & Oren 1995, regarding the predictive validity of the Psychometric Entrance Test in Israel), politicians, journalists and academicians keep pointing to these tests' (supposed) deficient validity. Despite several studies reporting the fairness (and even positive bias) of an admissions test towards minority, lingual and socially disadvantaged groups (e.g. Turvall, Bronner, Kennet-Cohen, & Oren, 2008), the prevailing public opinion is that the tests are biased against them. Coaching to the test is perceived as a threat to its validity, although there is empirical evidence to the contrary (Allalouf & Ben-Shakhar, 1998), and the content of the test is seen as irrelevant to higher education, though – again – the data proves that this is not the case.

In our opinion, the debate between the various branches of government – the executive, the legislature and the judiciary – on the one hand, and the universities on the other, although marred by gross misconceptions (on both sides), in the end serves to elucidate the issues, to inform the public, to improve the admissions process and to give voice to various social groups. It is thus another example of the strong connection between testing and social issues.

References

- Allalouf, A. & Ben-Shakhar, G. (1998). The Effect of Coaching on the Predictive Validity of Scholastic Aptitude Tests. Journal of Educational Measurement, 35(1), 31-47
- Balf, T. The Story Behind the SAT Overhaul. NYTimes Magtazine. March 6, 2014.
- Kennet-Cohen, T., Bronner, S. & Oren, C. (1995). A Meta-Analysis of the Predictive Validity of the Selection Process to Universities in Israel. RR 202. Jerusalem: NITE.
- Turvall, E., Bronner, S., Kennet-Cohen, T., & Oren, C. (2008). Fairness in the Higher Education Admissions Procedure: The Psychometric Entrance Test In Arabic. RR 349. Jerusalem: NITE.

54. Student Conceptions of Understanding and of Assessment Supporting Learning for Understanding

Rebecca Hamer (International Baccalaureate, The Netherlands) and Erik Jan van Rossum (Independent researcher, The Netherlands)

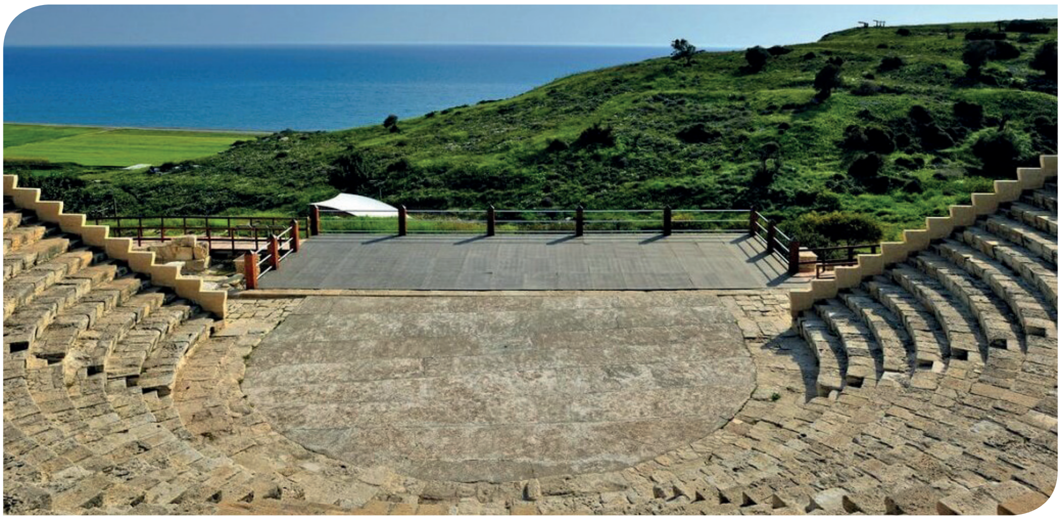
Rephrasing Perkins´ famous quote (1993), this paper addresses the question 'At the heart of assessment of understanding lies the very basic question: What is understanding?' Understanding is used in educational contexts and assessment often without further explanation, assuming that the meaning is self-evident and shared by students, teachers and researchers of learning alike. On a daily basis however, teachers are confronted with differences in how students interpret understanding affecting what and how they learn and know. Conversely, teachers' own interpretation of understanding shapes their teaching and so how students learn (for an overview see van Rossum & Hamer 2010). Lack of awareness of the existing range of conceptions of understanding and how these may undermine effective learning, teaching and valid assessment of learning should be a concern, but seldom seems to be. The current mainstream of research on understanding points appears to have not progressed beyond understanding as flexible performance as used within the Teaching for Understanding approach developed within Project Zero (Perkins, 1993). The learners´ focus on study success makes assessment a potentially powerful tool to influence learning (e.g. Gibbs, 1999) especially if assessment is geared towards awarding credit for deep understanding (Newble & Clarke, 1986).

The current paper presents a short review of the literature on conceptions of understanding up to 2015, linking it to epistemological models (e.g. Kegan, 1994) and providing an overview of twelve possible indicators to guide further research into more complex ways of understanding (e.g. Entwistle & Peterson, 2004; Barnett, 2004; Kegan, 1994). These indicators include the return of emotion in learning and deep understanding, connectedness and coherence, flexible adaptation to unknown situations, paradoxical thinking, alertness and open mindedness and identity development. It presents empirical qualitative data (N=167) supporting the recognition of at least two more complex ways of understanding that undergraduate students in higher education entertain, namely understanding-in-relativity and understanding-in-supercomplexity (Hamer & Van Rossum, in press; Barnett, 2004), as well as how students think such more complex understanding may be assessed. In the data students further discuss their views on existing assessment methods and their validity and success regarding assessing deep understanding, directly linking assessment methods to learning strategies and motivation to study.

It proposes the development of two matching approaches to curriculum development and aligned assessment based on the constructive alignment approach (Biggs, 2003), relativist alignment to encourage understanding-in-relativity and inner-alignment supporting the development of understanding-in-supercomplexity (Hamer & van Rossum, in press).

References

- Biggs, J.B. (2003). Teaching for quality learning at university: what the student does. Second edition. Berkshire: SRHE and Open University Press.
- Gibbs, G. (1999). Using Assessment Strategically to Change the Way Students Learn. In: Brown, S. & Glaser, A. (Eds.) Assessment Matters in Higher Education. (pp. 41-53).Buckinham: Open University Press.
- Entwistle, N.J., & Peterson, E.R. (2004). Conceptions of learning and knowledge in higher education: Relationships with study behaviour and influences of learning environments. International Journal of Educational Research, 41, 407-428.
- Kegan, R. (1994). In over our Heads. – The mental demands of modern life. Cambridge, Mass: Harvard University Press.
- Newble, D.I., & Clarke, R.M. (1986). The approaches to learning of students in a traditional and an innovative problem-based medical school, Medical Education,20, 267-273.
- Perkins, D. (1993). Teaching for Understanding. American Educator: The Professional Journal of the American Federation of Teachers, 17 (3), 28-35. Document available online http://www.exploratorium.edu/IFI/resources/workshops/teachingforunderstanding.html, 24 July 2011.
- Van Rossum, E.J., & Hamer, R. (2010). The Meaning of Learning and Knowing. Rotterdam: Sense Publishers.
- Hamer, R. & Van Rossum, E.J. (in press). Students' Conceptions of Understanding and Its Assessment. In E. Cano & G. Ion (Eds.), Innovative Practices For Higher Education Assessment and Measurement. Hershey, PA: IGI Global.



Session S: Assessment policy reforms

55. Changes in school accountability measures – Is there an effect on Enquiries about Results behaviour?

Sara Humphries, Vikas Dhawan and Beth Black (Ofqual, United Kingdom)

Schools in England are subject to high-stakes accountability measures which place a large amount of pressure on students achieving good grades at GCSE (General Certificate of Education – the most commonly taken qualification by 16 year olds). This pressure can lead to unintended consequences, attributed to a focus on stakeholders' reactions rather than regard for students' learning (Altrichter & Kemethofer, 2015). Currently, the headline accountability measure reported in school league tables is the percentage of candidates receiving the top four grades, i.e. grade C and above, with schools achieving below 40% open to increased scrutiny and potential closure. A concerning consequence of this is a tendency for schools to focus upon C/D borderline students, potentially at the expense of lower and higher achievers (Hutchings, 2015). One resulting behaviour of this focus is the tendency for schools to submit a disproportionate number of EaR (Enquiry about Results) challenges at the D/C borderline (Ofqual, 2016).

As of 2016, changes in government policy will see this headline accountability measure change from a single threshold target to a 'value added' model. 'Progress 8' is calculated as the progress made by students from the end of primary school achievement to GCSE (across all grades) in eight subjects. These changes could potentially influence where schools aim their resources, seeing a possible move away from the attention on C/D borderline students.

This paper discusses the findings of a study that explored the potential impact of these policy changes in accountability on the EaR system, through which candidates can appeal for exam boards to review a grade if they believe the original result is a product of marking error. For the 2015 summer exams, approximately 300 schools volunteered to pilot the new 'Progress 8' measure a year early. Student-level EaR data for the 2014 and 2015 exam periods were obtained from the exam boards and, in addition to the 'pilot' schools, a statistically comparable group of non-pilot schools was identified to act as a control. An investigation into the grade positions that these schools targeted their EaRs before and after the change will enable an understanding of the impacts that recent government policy changes may have in future years.

An EaR challenge that does not result in a grade change comes at a financial cost to the student or school, raising concerns that those candidates who can afford to pay or who are attending schools that can afford to submit more EaRs are at an unfair advantage, threatening educational and social equality. The data collected in this study will explore the question of whether schools opting into Progress 8 have changed their EaR behaviour due to the change in government policy – perhaps submitting fewer challenges at grade D, but perhaps more at other grades, and give an indication of the ways in which the new accountabilities measures will impact on the UK assessment system.

References

- Altrichter, H., & Kemethofer, D. (2015). Does accountability pressure through school inspections promote school improvement? School Effectiveness and School Improvement, 26(1), 32–56.
- Hutchings, M. (2015). Exam factories? The impact of accountability measures on children and young people. London: National Union of Teachers.
- Ofqual. (2016). Enquiries about results for GCSE and A level : Summer 2015 exam series.

56. Assessment and policy discord: transition to secondary level education in Northern Ireland

Leanne Henderson (Queen's University, United Kingdom)

Since the end of regulated transfer testing at age 11 in Northern Ireland (2009) statutory guidance issued by the Department of Education for Northern Ireland (DENI) for transfer to secondary level education states that 'Decisions on admission to post-primary [secondary] schools should not be based on the perceived academic ability of an applicant however defined or assessed' (DENI, 2015). Whilst, at present, there is no statutory arrangement for academic selection at eleven, The Northern Ireland (St Andrews Agreement) Act, 2006, represents a statutory provision by recognising, in law, the status of academically selective schools as a separate school type. This tension between statutory guidance and legal provision can be best described as policy discord: where two seemingly incompatible policies are operating concurrently. Statutory guidance (DENI, op. cit.) with regard to academic selection is, in practice, optional for some secondary schools. Currently 31% of secondary schools consider pupil outcomes from two commercial, and thus unregulated, transfer tests to inform admissions.

This paper reflects on research into the views and experiences of children as they navigate an assessment system, based on the use of two unregulated tests, which has emerged in this context of policy discord. The presentation will outline the purposes, processes, and outcomes of an online survey of transition age children (10/11/12 years old). The survey was developed in collaboration with a children's research advisory group (Lundy & McEvoy, 2012) made up of children in the first year of secondary school who had recent experience of transfer, and an appropriate level of expertise around the substantive issues of the research. Three key themes formed the basis of data collection: the current arrangements for transfer; the lack of transparency around the procedures; and the impact of the unregulated tests on children's learning in the final years of primary school (Elwood, 2013). These themes, whilst not particular to transition in Northern Ireland, are complicated by the context of policy discord and unregulated testing arrangements.

The survey collected qualitative and quantitative data, to address gaps in existing knowledge, about children's views and experiences as they transition to secondary school within a complex system of statutory and non-statutory arrangements. The data provides information about the assessments taken by children, including their self-reported outcomes, and their understanding of how their place at secondary school was awarded. This is particularly important because the tests are unregulated, therefore there is no duty on schools, or test providers, to make this information publicly available. The findings develop our understanding of: how the tests affect children's experiences of upper primary school; how the test outcomes are being used within admissions; and how their use contributes to lengthy, complex and uncertain admissions arrangements.

The transition to secondary level education is a significant landmark in a child's school career but the current policy discord around academic selection means that children navigate a system with additional complexities and uncertainties. For this reason it is particularly important to place their views at the centre of research which can be used to engage with policy makers who have a duty to give those views due consideration (United Nations, 1989).

References

- DENI, 2015. Post-Primary Transfer Policy, Bangor: DENI.
- Elwood, J., 2013. Educational assessment policy and practice: a matter of ethics. Assessment in Education: Principles, Policy and Practice, 20(2), pp. 205-220.
- Lundy, L. & McEvoy, L., 2012. Children's Rights and Research Processes: Assisting children to (in)formed views. Childhood, 19(1), pp. 129-144.
- Northern Ireland (St Andrews Agreement) Act, 2006. C.53, Section 21: Amendment of Education (Northern Ireland) Order 2006 etc. HMSO: London.

- United Nations, 1989. United Nations Convention on the Rights of the Child, Geneva: United Nations.

57. Reforming educational assessment in Trinidad and Tobago

Bas T. Hemker, Cor Sluijter (Cito, The Netherlands) and Newman Burdett (Independent, United Kingdom)

As part of a larger project in Trinidad and Tobago dedicated to Seamless Education, Cito and NFER were asked to assist the Ministry of Education in reforming educational assessment. This related to all aspects of examinations, testing and assessment (ETA), meaning that the ETA system, ETA quality, ETA production process and the organisation of the ETA were all considered. In order to support with the seamless education, item response theory (IRT) models and programs were introduced and explored. These techniques make it possible to track student's ability over time during the many stages in their educational career.

In the ETA system the interrelations of the various tests, ranging from early childhood education to pre-university exams were evaluated. What tests are being used for what purposes?
The quality of the tests was considered with an evaluation framework developed to be fit for purpose in the local situation. The production process referred to the steps involved in the production of test, which was highly related to the organisation: the roles and responsibilities of those involved in the test production.

In the presentation the focus is on the ETA system: the role of testing in the educational framework in Trinidad and Tobago. This started with a document analysis and interactive investigation. An important conclusion was that all educational assessment was considered to be high stakes even if there was no summative purpose of the test. This was true for all stakeholders, such as children, parents, teachers, principals, school boards and politicians, while at the same time they were uncomfortable with this reality. As a result formative testing was not used to the full extent. Tests were considered as tools to judge the students, teachers and school, rather than to help students in their learning process.

The main challenge was to design an ETA framework that involved assessment for learning, and to have a good distinctions between the formative tests and the summative tests. As the goal was to have assessment in the framework of Seamless Education, a standardized student monitoring system (SMS) was designed with a formative function. In order to guarantee this formative function a stakeholder review was implemented. The main questions here were to determine for all stakeholders how they could benefit from the SMS for them, how they could possibly misuse the SMS, and how to mitigate this misuse. Based on this, the design of the SMS was not only based on content and methodological criteria, but also involved an implementation strategy to help with the shift of the assessment paradigm.

During the project, many important social and political issues needed to be tackled. Issues such as the public trust in high-stakes assessments, the trust in teacher assessment and the impact of the assessment on the reproduction of social stratification. The shift in assessment paradigm towards more formative testing, the tension between assessment for learning and accountability was an important issue that only could be solved by the involvement of all stakeholders.
The development of the SMS was also supported by the introduction of IRT models into the production of the test, which means that technological and psychometric innovations in assessment had an impact here.

The experiences during this project helped to look afresh at educational assessment in The Netherlands. The presentation will conclude with the main learning points from this experience that are useful in practice in The Netherlands as well.

References

- Hemker, B.T., Sluijter, C., Sharp, S. and Burdett, N. (2015). Final Report – Support for a Seamless Education System with regard to Examination, Testing and Assessment. Report 12. Cito, Arnhem.

Session T: Talking about the fundamental issues when validating assessments

58. Validity arguments for considering different expectations when setting cut-off scores in a formative assessment in digital responsibility

Ingrid Radtke (Vox – Norwegian agency for lifelong learning, Norway) and Ove Hatlevik (University of Oslo, Norway)

The Norwegian Directorate for teaching and learning has introduced learning supportive assessments, a type of formative assessment, in basic skills in different subjects. The aim of these assessments is to inform teachers about the mastery level of their students. Cut-off scores are established because the results are used to give feedback to the students and to customise training in the classroom. Therefore, the teachers are provided with guidance material and access to all test tasks.

To decide about performance standards and the cut-off scores is a process of negotiation between different groups, for example stakeholders, experts and teachers. However, how are these decisions reached and which arguments are most trustworthy? Many standard setting methods follow the same procedures where a group of experts after a longer process decides about the final cut-score. Though little work has been done yet to monitor the content of the group discussions and to evaluate the external and consequential validity. In addition, recent research indicates that these methods provide us with different results about cut-off scores.

Pant, Rupp, Tiffin-Richards and Köller (2009) present validity issues in relation to standard settings procedures. They distinguish between procedural, internal, external and consequential elements. The two latter seem to be important when dealing with different results from standard setting methods. External validity evidence is about the comparison between different methods for setting cut-off scores. Consequential validity evidence refers to reasonable cut-off scores and whether they are perceived to be in line with the performance standards.

In the national curriculum, digital responsibility is defined as a sub-category of students' ability to use ICT. In 2016 a formative 50-item assessment in digital responsibility is available for Norwegian eight graders. The purpose of this test is to inform schools and teachers about the proficiency levels of the students in digital responsibility and to customise teaching and learning processes.

The digital responsibility assessment was administrated on a national sample of 1,026 students from 26 schools. An expert group of ten persons met for a one-day standard setting workshop. Two test-centred standard setting methods (Angoff and Bookmark) were used, and they gave different results for the cut-off scores. The question was then how to deal with these differences? One solution was to combine test-centred and students-centred methods, which is something that Cizek (2012) regards as a natural part of any standard setting.

This paper addresses external and consequential validity of standard-settings. We are looking into how four test-centred and student-centred arguments can be explicit used in the process of decision-making. First, as this is a voluntary test the role of the teachers is important since they have regard the proficiency levels of the test to be trustworthy. Second, the number of students on each of the levels plays a role, because there are expectations about a large group in the middle and 15-25 percentages of students at the upper and lower levels. Third, the number of items qualifying for a certain level is important. This is about realistic cut-off scores, and that cut-off scores are in alignment with the number of correct items. Fourth, the items located around the cut score has to be discussed because one could assume all items are related to one of the performance levels.

References

- Cizek, Gregory J. (2012): An Introduction to Contemporary Standard Setting: Concepts, Characteristics, and Contexts. In: Cizek, Gregory J. (Ed.), (2012): Setting Performance Standards: Concepts, Methods and Perspectives. Mahwah, NJ: Lawrence Erlbaum, pp. 3 – 14.
- Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P. & Köller, O. (2009). Validity issues in standard-setting studies. Studies in Educational Evaluation, 35, pp. 95–101.

59. Validation of student selection system in Kazakhstan: transparency and accountability

Miras Baimyrza, Zamira Rakhymbayeva (Nazarbayev Intellectual Schools, Kazakhstan), Caroline Jongkamp and Frans Kleintjes (Cito, The Netherlands)

Nazarbayev Intellectual Schools (NIS) has developed new system of student selection for 7th grade who can be taught math and science oriented programs at 20 NIS throughout Kazakhstan. For the first time the new system was administered in 2013. Since then more than 45 thousand candidates participated in the test. More than 8 thousand therefrom were selected to NIS.

In the paper we will focus on the longitudinal research of predictive validity of the student selection system. For the analysis purpose, quantitative data of students' who were selected starting from 2013 were collected. As quantitative data test results of subject test (Mathematics, Kazakh language, Russian language, English language) and CTY ability test (Quantitative and Spatial reasoning) of selection test; as external criterion student school achievement results were used: formative and internal summative assessment results of scientific, humanitarian and language subjects were used.

In the political context of this high-stakes assessment, accountability and transparency are of crucial importance to attain both public and political support. It will be described how the transparency is achieved by providing adequate information to the stakeholders through various communication channels. Accountability is supported by validation research.

The key question for the validation research is: have the right students been selected in this Selection procedure? The results from LISREL analysis [Jöreskog & Sorbom, 1996] show that Selection test serves well in predicting success in grade 7 and 8 performance in Mathematics, and moderate in predicting performance in languages. The results from a latent class analysis [Lazarsfeld & Henry, 1968] show that selected students can be clustered in three types according to their abilities in different subjects. This information raises new debates for policy makers.

The paper describes the interaction between policy, public trust and validation in high stakes assessment in Kazakhstan.

References

- Jöreskog, K. G., & Sorbom, D. (1996). LISREL 8: User's reference guide. Chicago, IL: Scientific Software International.
- Lazarsfeld PF, Henry NW. (1968). Latent structure analysis. Boston: Houghton Mifflin.

60. Helping the industry to engage with validity and validation

Paul Newton (Ofqual, United Kingdom)

In 2014, Ofqual changed its approach to regulating the assessment industry in England, by putting validity at the heart of what it does. Although this innovation was welcomed by the industry, it raised questions concerning exactly what Ofqual meant by validity and how it might expect awarding organisations to demonstrate this, through validation.

The meaning that Ofqual had in mind was similar to the modern ‘unitary’ view promoted by Samuel Messick (1989) in his classic treatise on validity. Unfortunately, the modern view is far from accessible; and even the argument-based approach to validation, which was developed in an explicit attempt to make validation more accessible, has achieved only limited success in recent years.

After identifying problems with the way in which the modern view of validity is typically described, and with the argument-based approach to validation, the presentation will outline steps that Ofqual has taken to help the assessment industry in England to engage with validity and validation, by developing an alternative approach to thinking and talking about these fundamental ideas.

The foundation for this work was a ‘common sense’ view of validity which asks whether we are assessing the right thing in the right way to produce accurate and useful assessment results. We formalised this – to provide a technical point of reference for subsequent discussion and clarification – by stating that validity is the degree to which it is possible to measure what needs to be measured by implementing an assessment procedure. Our intention was to signal to the assessment industry that validity is a technical concept related to educational measurement;

whilst also signalling that measurement is not an end in itself but a means for supporting wider purposes, and that those wider purposes dictate what needs to be measured.

The identification of a paradox in the way in which recent editions of the Standards for Educational and Psychological Testing (e.g. AERA, APA and NCME, 2014) characterise validation evidence led us to distinguish between two quite different 'lenses' through which to scrutinise the validity of assessment procedures. The first, macro-validation, involves evaluating the overall macro-validity claim (that it is possible to measure what needs to be measured) on the basis of measurement outcomes and systemic impact. The second, micro-validation, involves evaluating underpinning micro-validity claims concerning the effective design of features and processes which comprise the assessment procedure. Interestingly, the history of ideas on educational assessment validation can be portrayed as a gradual rejection of the macro-validation mind-set which dominated early thinking.

Finally, in response to problems associated with argument-based validation frameworks, we developed a design-based validation framework, using the structure provided by the educational assessment 'lifecycle'. This paid dividends in terms of both accessibility and comprehensiveness, providing practitioners with a more intuitive approach to gathering and organising a full range of validation evidence and argument.

Validity is the principal underpinning concept for educational assessment; and, in the context of industry regulation, its social and political significance is foregrounded. This presentation is therefore highly relevant to the theme of the conference and links directly to a number of the 'possible topics' identified by the Programme Committee: regulation of the assessment industry; the social responsibility of Examination Boards and Awarding Bodies; validity issues in educational assessment.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. Washington, D.C.: American Educational Research Association.
- Messick, S. (1989). Validity. In R. Linn (Ed.). Educational Measurement (3rd edition) (pp.13-100). Washington, D.C.: American Council on Education.

Session U: Supporting teachers in assessment

61. Computer-based formative assessment in classrooms

Stéphanie Berger (University of Zurich, Institute for Educational Evaluation/ University of Twente, Research Center for Examination and Certification, Switzerland), Urs Moser (University of Zurich, Institute for Educational Evaluation, Switzerland), Angela Verschoor (Cito, The Netherlands) and Theo J.H.M. Eggen (Cito/University of Twente, Research Center for Examination and Certification, The Netherlands)

Modern computer technologies have fostered the development of software for formative classroom assessment. One obvious advantage of computer based assessments over paper-based assessments is that they allow for automated scoring and immediate reporting. Thus, teachers do not need to invest time in the scoring of assessment items, but they can concentrate on the interpretation of the assessment outcomes and on the preparation of aligned teaching activities (Hattie & Brown, 2007). Besides, software for computer-based assessment can also assist teachers in efficiently creating objective and reliable assessments for individual students.

In Northwestern Switzerland, an online item bank for formative assessment has recently been introduced. The purpose of this online item bank is to provide teachers an instrument for data-based decision making, a specific approach of formative assessment (Van der Kleij, Vermeulen, Schildkamp & Eggen, 2015). This means, the online item bank is designed to supports teachers in efficiently assessing the strengths and weaknesses of their students in relation to the curriculum on demand during the school year. The assessment outcomes serve teachers as a starting point to plan the next steps for the students along their individual learning paths and to monitor the students' progress over time.

In this paper, we will elaborate how we combined item response theory methods, advanced computer technologies and the new Swiss curriculum to develop the online item bank and to provide teachers a flexible but at the same time also standardized instrument for formative assessment. Furthermore, we will discuss three different approaches to support teachers in creating assessments by means of computer software and refer to how we implemented this approaches in our online item bank for formative assessment. First – if teachers are interested in assessing specific competences, computer software can support them in identifying suitable assessment content (i.e., assessment items) by means of key words and filtering techniques. Therefore, we implemented an integrated item banking system in which we categorized all items based on their content according to the competence levels that are stated in the curriculum and added psychometric item parameters such as the item difficulty. The second approach focuses on a more general level of the assessment creation process: Assessment blueprints provide teachers guidelines to create reliable assessment by, for example, requesting a minimal test length or content balancing within a single assessment. Third, computer algorithms can be used to fully automate the item selection and to adaptively select items that correspond to the actual ability level of each individual student. Such computer adaptive tests (CAT, e.g., van der Linden & Glas 2010) provide more reliable measurements of the ability of each single student than general linear assessments.

In addition to the methodological discussion of the three approaches, we will present some first experiences from the implementation of the system in classrooms.

References

- Hattie, J. A. C., & Brown, G. T. L. (2007). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. Journal of Educational Technology Systems, 36(2), 189–201.
- Van der Kleij, Fabienne M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, Assessment for Learning and diagnostic testing in formative assessment. Assessment in Education: Principles, Policy & Practice. (ahead-of-print), 1–20.
- van der Linden, W. J., & Glas, C. A. W. (2010). Elements of adaptive testing. New York: Springer.

62. Our TALE: the importance of the social and educational context of assessment

Dina Tsagari (University of Cyprus, Cyprus), Karin Vogt (University of Heidelberg, Germany), Ildiko Csepes (University of Debrecen, Hungary), Tony Green (University of Bedfordshire, United Kingdom) and Nicos Sifakis (Hellenic Open University, Greece)

Educational assessment policies and programmes, operating within various contexts and supported by national or international agencies, have long been the focus of discussion and research in the field of educational assessment. Nevertheless, the last decades have witnessed varying assessment initiatives at the classroom-based assessment level that have placed increasing emphasis on the professional development of teachers. For example, within the field of language testing and assessment (LTA), the growth in the use of accountability systems and the influence of external frameworks in educational policy making, such as the Common

European Framework of Reference for Languages (Council of Europe, 2001), have increased both the amount of LTA required of English Language Teachers (ELTs) and the importance placed on it. ELTs are now expected to design and score a wide variety of language tests that are relevant to their own particular teaching context; carry out innovative assessment procedures (self- and peer-assessment, portfolios, etc); provide useful feedback to learners based on results of such assessments; and align their LTA methods with language curricula or educational policies in ways that meet national or European language assessment standards. To be able to do so, teachers need to reach and acquire high levels of 'language assessment literacy' (LAL), that is skills, knowledge, methods and techniques needed to design and carry out effective assessment tasks and to make informed decisions based on assessment data (Fulcher, 2012; Harding & Kremmel, 2016; Taylor, 2013).

However, research has shown that in many educational systems across Europe, ELTs are not able to create good quality assessment materials and procedures. This is at least partly because they are not sufficiently trained in LAL (Hasselgreen et al. 2004; Tsagari & Csepes, 2011; Vogt & Tsagari, 2014). Also research has pointed to the fact that assessment literacy is not a straightforward matter. Teachers' acquisition and implementation of LAL seems to be a situated activity, located in particular contexts, each characterized by assessment practices compatible with the social and educational values and beliefs that the 'school' community upholds, hence the school's assessment culture (Inbar-Lourie, 2008; 2013). Recently, studies have begun looking at LAL in particular contexts drawing attention to the intricacies in examining teachers' perceptions and knowledge about the learning and assessment they bring to the dynamic LAL acquisition process (Scarino, 2013; Xu, 2015).

The goal of this paper is to examine the notion of LAL on a constructive and interpretive epistemological basis taking into consideration the importance of 'context'. More specifically, the paper will focus on the first phase of the of a three-year European-funded project entitled 'Teachers' Assessment Literacy Enhancement (TALE)'. The first phase involved a needs analysis achieved through extensive consultation with ELTs and their students aiming to ensure that the future online LTA training course designed for the purposes of the project would meet their needs. The study adopted an exploratory method design based on quantitative data collected via online questionnaires designed to investigate how teachers perceive and practice assessment, the types of assessment they use, the impact of these assessments, teachers' training needs and modes of delivery of training courses. The survey was conducted among 829 ELTs in four European contexts identified in this project. Descriptive and inferential statistics have been used in order to examine the trends identified per country and also interrelationships between the LTA needs of ELTs in the various contexts involved in this project. The findings provide nuanced in-depth understanding of the assessment needs of ELTs and their students which contribute to the identification of assessment priorities and the development of assessment training strategies that are contextually situated.

63. Integration of standards for expected writing proficiency within an AfL-approach

Ragnar Thygesen (University of Agder, Norway), Lars S. Evensen and Gustaf B. U. Skar (Norwegian University of Science and Technology, Norway)

Low-stake models for every-day classroom assessment, developed on a theoretical and empirical basis, are warranted. This paper presents a model that includes a theoretical construct and local integration of novel standards for expected proficiency in writing as a means of promoting assessment for learning. The construct, The Wheel of Writing, is a functional model where acts of writing and corresponding purposes constrain the use of semiotic resources. Empirical collaboration with teachers was undertaken, defining which specified criteria for judging writing competency could reasonably be set at the start of grades 5 and 8. Criteria were identified in various domains of writing: Communication, Content, Composition, Use of language and grammar, Orthography, Punctuation, Multimodal presentation. The set of criteria represents the standards.

AfL in low-stakes contexts raises problematic issues related to standards development.
The presentation discusses how a shared rhetorical community among teachers can develop over time that will produce reliable assessment of students' texts across local contexts when standards are integrated within an Afl-approach.

The study reported was designed as a repeated measurements field study over a three years span, with 20 intervention and 5 comparison schools. Students' texts and teachers' assessments of texts were gathered as pre/posttest data, plus through 4 measurements in between.
The training program consisted of instruction to teachers with a focus on construct, assessment domains, writing standards, and principles for developing assignments. A rubric sheet was used along with guidelines for assessment. For each domain, texts were thus assessed according to the standards and scored with use of a five-band ordinal scale.

Texts were read by four teachers (two rater pairs), which enabled monitoring of the overall rater agreement using estimates of the intra-class correlation coefficient, one-way single measure. Three sets of reliability estimates were computed for 3rd-4th grade and 6th-7th grade teachers. The first estimate summed assessment across all domains. The second summed assessment of 'functional' domains: communication, content, text structure and language use. The third summed 'coding' domains: orthography and punctuation.

Results indicate that reliability in general increased over time. However, estimates of reliability for assessment of functional and coding domains show that major improvement was related to assessment of functional domains only. For coding domains the results demonstrate a surprising decrease in reliability. It may seem that a strong interpretive community had been present in this area at the start of the project.

Changing local communities of writing into a joint rhetorical community proved difficult, primarily due to time: The pre-existence of stronger existing local communities of assessment practice than expected, where focus was on one dimensional aspect of writing only (orthography and punctuation) interfered. Adapting a wider construct proved time-consuming and interfered with teachers' expertise in applying their previous construct. Also, assessing only the text at hand in relation to multidimensional standards turned out to be a qualitatively different approach for many. Making such a move resulted in temporarily poorer reliability. Accordingly, 'sub-surface' validity development may in fact be indicated by temporarily poorer reliability.

Consequently, validity concerns have to be given priority before reliability concerns are given equal priority whenever a new construct is part of an intervention. Forging a rhetorical joint community for assessment out of several local communities of practice require prolonged investment.

References

- Berge, K.-L.; Evensen, L. S. & Thygesen, R. (2016). The Wheel of Writing: a model of the writing domain for the teaching and assessing of writing as a key competency. The Curriculum Journal. Published online 21 January.
- Evensen, L. S.; Berge, K.-L.; Thygesen, R.; Matre, S. & Solheim, R. (2016). Standards as a tool for teaching and assessing cross-curricular writing. The Curriculuar Journal. Published online 21 January.

Session V: Scales, scores, analysis

64. TIA-Excel – an easy tool for test and item analysis in classroom assessments

Eef Ameel and Rianne Janssen (KU Leuven, Belgium)

The importance of Test and Item Analysis
At the level of every-day classroom assessment, more countries are focusing on assessment carried out by the teachers themselves. To evaluate and to guarantee the psychometric quality of classroom assessments, Test and Item Analysis (TIA) is a useful process, though rarely used in the context of classroom assessment. TIA examines student responses to individual test items in order to assess the quality of those items and of the test as a whole. Besides insight in the quality of the test, TIA also provides teachers with useful information about the performance of their students at the level of individual items. The current paper presents a new spreadsheet application in Excel, called TIA-excel, that can be easily employed by teachers to screen the quality of a test with dichotomously scored items or multiple choice items.

Item and test characteristics in TIA-excel
In TIA-excel a number of item and test statistics are reported which aid in evaluating the effectiveness of an item and the quality of the test as a whole. For each item, the difficulty (p-value) is computed by the proportion of students choosing the correct response. For each wrong answer alternative, the degree of 'attractiveness' can be derived from the proportion of students who chose the respective alternative (= a-value). Further, item discrimination is measured by different indices (e.g., standard deviation, item-total correlation,...). Also reliability estimates are included (e.g., Cronbach's alpha). Looking at all these item and test characteristics will assist the test developer (teacher) in determining what is good or wrong with individual items.

Additional visualization tool in TIA-excel
A visualization tool was integrated in TIA-excel, based on score group analysis (Veldhuijzen, Goldebeld, & Sanders, 1993). Students are divided into different score groups, depending on their total test score. For each score group, the p- and a-values are graphically represented per item. This plot can be seen as an empirical counterpart of an item characteristic curve in item response theory. The included visualization tool in TIA-excel allows teachers to gain immediate insight in the quality of the items and the test as a whole.

Low-cost, easy accessible and transparent tool
To perform a TIA, many software and feedback techniques have been developed. Compared to other tools, there are several advantages of TIA-excel: first, it is low-cost, since almost everyone has access to Excel. Second, TIA-excel is an easy accessible tool, also for non-statisticians. Third, the visualization tool helps the test developer to derive the quality of the test and her individual items at first glance.

Relation to the conference theme
At the level of every-day classroom assessment, more countries are focusing on assessment carried out by the teachers themselves. Test and item analysis, even though it is rarely applied, is very useful for teachers to evaluate the quality of tests. TIA-excel was especially developed to make test item analysis easily accessible for teachers. As TIA-excel helps to increase the quality of classroom assessments, it may contribute to the professional development of teachers as examiners and may lead to an increased trust in teacher assessment.

References

- Veldhuijzen, N. H., Goldebeld, P., & Sanders, P. F. (1993). Klassieke testtheorie en generaliseerbaarheidstheorie. In T. Eggen & P. F. Sanders (Eds.), Psychometrie in de praktijk (pp. 33-85). Arnhem: CITO.

65. Elimination scoring as an alternative for correction for guessing in multiple-choice questions: an empirical comparison

Rianne Janssen, Jef Vanderoost and Tinne De Laet (KU Leuven, Belgium)

Introduction

In universities high-stakes exams are commonly conducted using multiple-choice items given their ease of administration and scoring. A disadvantage of multiple-choice items is that a correct answer can also be obtained by guessing. In order to discourage guessing, multiple-choice questions are frequently scored using the so-called 'correction for guessing', which gives a penalty to wrong answers equal to 1/(k -1), with k being the number of response alternatives. Examinees are still advised to guess whenever they can eliminate at least one response alternative. However, there are alleged differences among examinees with respect to their willingness to guess. Studies have also shown that girls are less eager to guess.

An alternative way to score multiple-choice questions is using elimination scoring: examinees are asked to eliminate all response alternatives they think are incorrect. When they can eliminate k – 1 alternatives, their response corresponds to choosing one alternative. When students can eliminate less than k – 1 alternatives, they are able to show their partial knowledge without the need to guess the answer.

An empirical study

The traditional way of scoring multiple-choice questions using correction for guessing was compared with elimination scoring using a within-subject design.

Method

145 students at a large Flemish university sat a try-out exam halfway a course on the history of psychology. The exam consisted of 20 multiple-choice questions with four response alternatives each. All students answered the first set of 10 items with the traditional instructions and correction for guessing and the second set of 10 items with elimination scoring. There were two test booklets randomly distributed to the students. These booklets contained the same items, but the position of the two sets of 10 items was switched, allowing to compare the two types of instructions by controlling for item content. In both scoring methods, the reward for a fully correct response and for a wrong response (choosing a wrong alternative or eliminating the correct response) were equal. In elimination scoring partial credit was given for correctly eliminating less than k–1 response alternatives. After the exam, the students were given a short questionnaire on their opinions about the two answering instructions.

Four weeks later (on May 9, 2016), the students took The Alper and Haber Achievement Anxiety Test (1960), which is a questionnaire that assesses anxiety related to exams as either inhibiting or enhancing performance.

Results

- Students performed rather lowly on the low-stakes exam. The mean result for elimination scoring was 2.81 (SD = 2.00) and 3.15 for correction for guessing (SD = 2.30). However, there were significant interactions between the two scoring methods at the item level.
- Examinees agreed that the elimination instructions were clear but that these required more time than the tradition instructions. Students differed in opinion as to whether elimination scoring would reduce their exam stress. Overall, a slight majority preferred elimination scoring.
- The additional questionnaires on test anxiety and guessing behavior in multiple-choice questions will be correlated with (a) differences in performance on the two scoring instructions and (b) examinees' preference for the scoring system.

Conclusions

The results of the present study suggest that elimination scoring is a promising alternative for correction for guessing. However, the study also shows that a change in instructions should be accompanied by giving extensive information and training to the examinees.

Relation to the conference theme
The discussion how to score multiple-choice exams has been a topic of public debate among Flemish universities with political agenda's interfering with psychometric considerations.
The present study illustrates evidence-based policy making in assessments.

66. Number identification: A scale identifying a progression pathway

Sarah Gott, Lee Copping, Helen Cramman, Christine Merrell (CEM, University of Durham, United Kingdom) and Peter Tymms (University of Durham, United Kingdom)

This study aimed to see if the results of an assessment of a child's ability to identify a number forms a developmental scale such that performance and progress may be measured using this one instrument. The focus of the analysis was to determine if this scale captured a single latent trait which proved to be reliable and invariant across a number of factors, and from which valid interpretations may be made about a child's learning trajectory.

Over 10,000 children in Reception class (aged 4) in England or Primary 1 in Scotland in the academic year 2012/13 completed the Performance Indicators in Primary Schools (PIPS) computerised monitoring system run by the Centre for Evaluation and Monitoring at Durham University, UK (see www.cem.org and Tymms (1999) for more information).

The full PIPS assessment includes sections which assess language, early reading and mathematics development. A child's ability to identify numbers was assessed on a one to one basis at the start and end of the reception year. The software presents questions verbally using audio files. For the section which assesses number identification, children see a number on-screen and are asked 'What is this number?' to which they respond verbally. The teacher records the pupils' answers on-screen as either right or wrong. Numbers are ordered in increasing difficulty, beginning with the most commonly recognised single digits and then presenting three randomly generated numbers in the teens, then three two-digit numbers and finally, five three-digit numbers. Each child is presented with a maximum of 21 items. The internal reliability (Cronbach's alpha) for this section of the assessment is 0.93 (Tymms et al., 2012).

The fit of the data to the Rasch measurement model was assessed. Item difficulties at the start and end of the academic year were established, and invariance of these estimates were investigated using methods suggested by Bond and Fox (2007) following on from work by Smith Jr (2001). This involved a means of establishing bias in the test key factors. Criteria for the size of the difference and its statistical significance were used.
Further methods proposed by Wright (2003) were used to investigate how much progress students showed across the first year at school.

In conclusion, it is proposed that a child's ability to identify numbers develops in a clear progression through an established order of single digit numbers, two digit numbers and finally three digit numbers. The analysis suggests that this scale is invariant across key factors. Our research makes a distinctive contribution to the current literature, demonstrating that the scale remains invariant across a year of teaching. This is discussed in the full paper and practical suggestions made for use in the classroom.

References

- Bond, T.G., & Fox, C.M. (2007). Applying the Rasch Model. Fundamental Measurement in the Human Sciences. (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Smith Jr, Everett V. (2001). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. Journal of applied measurement.
- Tymms, P. (1999). Baseline assessment, value-added and the prediction of reading. Journal of Research in Reading, 22(1), 27-36.

- Tymms, P, Merrell, C, Henderson, B, Albone, S, & Jones, P. (2012). Learning difficulties in the primary school years: Predictability from on-entry baseline assessment. Online Educational Research Journal (available at: www. oerj. org).
- Wright, B.D. (2003). Rack and Stack: Time 1 vs. Time 2 or Pre-Test vs. Post-Test. Rasch Measurement Transactions, 17(1), 905-906.

Session W: Methodological advances in assessment

67. High-stakes assessment instruments: one-size-fits-all versus flexibility

Caroline Jongkamp and Angela Verschoor (Cito, The Netherlands)

In the Netherlands, systems of digital, flexible examination have been introduced more than 10 years ago. Flexible examination means that a test can be administered at any time and any place. This flexible system is being applied in the Netherlands, in final exams for secondary education and in intermediate vocational education.

In this presentation, we will share how automated test assembly methods have proven to be useful to achieve flexibility in organizing the exams without adding to the risk of security breach. Instead of having one exam version for all students, the flexible system involves having many versions. Each of those variants can be employed at a different period of time, thus enabling candidates to take an exam when they feel ready for it. The versions should be different, but equivalent with respect to the statistical properties to classify test-takers on the same ability scale regardless of the version and timing of the test administration.
The challenge to assemble sufficient quantity and quality of test versions at acceptable costs, is successfully met by applying standard optimization techniques in the test assembly process. Van der Linden (2005) has shown how these standard linear optimization models can be employed in test assembly. Verschoor (2007) has formulated several methods to allow overlap between variants in such a way that security breaches can be maintained to a desired risk level while at the same time limiting item production to an acceptable level.

Two different models have been developed. In some examinations, pretesting of items in a relatively small sample of the population is common practice. The purpose of those pretests is to collect data regarding the items. Usually, those data are analyzed using Item Response Theory. The parameters, acquired in this way, are used in the test assembly models to ensure that all variants are equally difficult and have a high reliability. Thus, all candidates will have equal possibilities to show their potential, without need for advanced equation techniques.
At the same time, the variants have only limited overlap without any apparent structure, thus minimizing the risk of security breaches. Students taking their exams later in the period have no 'artificial' advantage over those taking their exams earlier.

In other circumstances, pretesting is not allowed or is infeasible. Here, a strict regime of re-use can be employed to assemble variants of equal difficulty. In final examinations in secondary education, some items are allowed to be re-used after two or more years, while all variants should contain at least 50 percent newly developed items. Thus, variants of exactly equal difficulty and reliability are not created, but over the years differences have shown to have roughly halved. In case all items can be re-used, for example in the entrance tests for teacher colleges, differences have been reduced to a negligible level.

Other models are in use outside the Netherlands. In the selection tests from NIS in Kazakhstan, seeding mechanisms are used. In each variant, a small number of newly developed items are used, while they do not count towards the test outcome. Thus, students are not punished if their variant proves to be more difficult than others.

It will be shown that the methods applied have a high degree of flexibility when it comes to adapting to particular socio-political contexts. As such, they are highly effective for high-stakes assessment systems where many stakeholders and complex expectations are involved.

References

- van der Linden, W.J. (2005). Linear models for optimal test design. New York: Springer.
- Verschoor , A.J. (2007) Genetic Algorithms for automated test assembly. PhD thesis Twente University, Enschede

68. Innovations in standardized testing in Lithuania. Measuring Higher order Thinking Skills in Lithuania

Eglé Melnike (NEC, National Examinations Centre Lithuania, Lithuania) and Frans Kleintjes (Cito, The Netherlands)

The National Examination Centre executes a project called 'Development of standardised student achievement assessment and self-assessment tools for general education schools' The project aims to address the problems stated in the strategic documents on education related to quality assurance in education, development of the learner-monitoring system, school self-evaluation, strengthening of the relationship between external and internal school evaluation, and curriculum modernisation. The project aims to implement in a comprehensive way the innovations in assessing student achievements, data analysis, and the use of the collected information for teaching/learning improvement.

During the project, a national study on students' achievements has been carried out and a more comprehensive and deeper use of standardized tests in schools are tested.
The following standardised programmes/testing standards for student achievement assessment were created: Grade 4: Lithuanian Language (reading and writing) and Mathematics and Grade 8: Lithuanian Language (reading and writing), Mathematics and History. Next work on standardised programmes/testing standards and standardised tests for the following subjects and grades were continued: Grade 4 – World Science; grade 8 – Natural Sciences and Social Sciences, a new development has been added to the History programme and covers Geography; and grade 6 – Lithuanian Language (reading and writing) and Mathematics. Early diagnostic assessment tools for assessing learning achievements of grade 2 students are also being developed. Their aim is to assess students' achievements and the existing achievement gaps after they graduate from grade 2 in order to provide assistance in overcoming any difficulties. The fundamental aim of early diagnostic assessment is prevention of learning difficulties, problems and failures. The tool for diagnostic assessment of grade 2 students will focus on achievements in the Lithuanian Language (reading and writing) and Mathematics.

There is no compulsory assessment of all students using standardised tests in Lithuania. It is up to a school or municipality to do so. The National Examination Centre only prepares the tools – standardised tests and instructions as to how to assess students' replies. Schools subsequently carry out testing and assessment. Additional automatic test result calculation is also offered to the schools. For this purpose schools must enter students' results in the electronic record form prepared by the National Examination Centre. In this case the school will receive automatically generated reports on several levels – the student, class and school level. The reports display the outcomes according to the achievement levels in a subject, area of the subject (e.g. in the case of Mathematics: numbers and calculations, algebra, geometry, etc.), and groups of abilities (e.g. knowledge and comprehension, application, higher-order thinking skills). Comparative information regarding the number of students in a certain level of achievements (satisfactory, basic, higher) in the country, school or a separate grade is also provided.

This presentation and paper reports how the ambitions have been realized and also reports on how the National Examination Center has overcome major challenges in realizing the project aims. Together with experts from Cito, the Dutch national institute for educational measurement, the following topics have been addressed.

- Assessment of higher-level skills and competence of ability to learn
- What are the specific characteristics of creating relevant, authentic, contextual, and integrated assessment tasks corresponding to modern requirements.
- Creation of assessment tasks for assessing higher-level thinking skills, selection of an authentic context, specifics of structured tasks.
- Creation of assessment tasks of higher-level thinking skills for assessing low, medium and high achievement students.
- Creation of integrated, interdisciplinary standardised test tasks (e.g. reading and science, mathematics and science, etc.) for primary school.
- Creation of integrated, interdisciplinary standardised Science and Social Science test tasks for basic school (focus – 8 grade).

69. Predicting item difficulty: methodological challenges and way forward

Yasmine El Masri, Jo-Anne Baird (Oxford University Centre for Educational Assessment, United Kingdom), Steve Ferrara (Pearson USA, USA) and Peter W. Foltz (University of Colorado, USA)

According to the Standards for Educational and Psychological Testing (2014), validity is the most fundamental consideration in evaluating tests. However, construct-irrelevant variance is a major threat for validity. Construct-irrelevant variance may arise from numerous sources, such as lack of clarity in instructions, lack of consistency in test preparation, or by item or test formats (including the use of bubble sheets) that might be unfamiliar to certain populations. However, although research has examined these issues in regard to high stakes test, this is not necessarily the case for low-stakes tests such as in international studies.

In countries such as the USA with large testing cultures, students take many multiple-choice high-stakes tests starting from elementary school. Due to the high-stakes nature of these tests, teachers typically expose their students to test taking strategies (TTS), such as whether answer changing should be used, or whether answers should be omitted when unknown. However, such instruction cannot necessarily generalize to students in other countries with low testing cultures, or with tests that mostly include open-ended questions. Therefore, this study serves as a starting point to determine whether background factors differentiate elementary school student’s TTS, in a first attempt to determine whether such differences could be a threat to a study’s validity. The main purpose of the current study was a) to determine variables that could explain part of the variation in student’s test-taking strategies, and b) to determine whether some student’s exposure to the TIMSS study has differentiated their use of TTS.

The sample of this study is composed of 291 fourth-graders in Cyprus (which has a low testing culture), of which 45.3% were male. The students were administered a 30-item questionnaire that measured frequency of using test-taking strategies, attitudes, and background questions, The data were analyzed with the use of descriptive and inferential statistics (e.g. MANOVAs). The results have shown that differences do exist in certain TTS among groups of students. For example, students who had participated in academic competitions, and thus had more testing experience, were least likely to use answer-changing. However, students who had participated in TIMSS reported using answer-changing more frequently in occasions where they had reread and better understood the questions. This study also found that the students whose native language was not Greek, or who had not participated in academic competitions had more negative attitudes towards multiple-choice tests.

In conclusion, this study found that despite the homogeneous student culture in Cypriot elementary schools, statistically significant differences were found in student’s attitudes towards multiple-choice tests, and in regard to the TTS of answer-changing. One would expect, that the differences might be larger when compared to students from other countries, as was found in a PISA study where the highest achieving countries were the ones with the least amount of omitted response rates (Gillmore, Longback & Poggio,2014).

The significance of this study lies in the finding that background factors are related to TTS which could increase the test error variance. Therefore, from a research perspective, these results strengthen the need for the replication of this study for between country comparisons, with the administration of achievement tests to measure the actual error variance and their effects on validity. From an educational perspective, these results also identify the importance of uniformly familiarizing students with the testing process and with TTS in order to minimize the effects of backgrounds variables that are not construct related.

References

- Author,(2014). Standards for educational and psychological testing. Washington DC: American Educational Research Association.
- Gillmore,S.,Longabach,T.,& Poggio,J.(2014). A new threat to validity: An examination of cultural discrepancies in omission rates on international assessments. Paper presented on the AERA conference, Philadelphia,PA.

Session X: Accountability in assessment

70. Shifting Emphases: qualifications, accountability and school improvement

Tim Oates and Sylvia Green (Cambridge Assessment, United Kingdom)

In England performance tables contain a large number of measures of attainment and of pupil progress. These measures have been changed over time, with this change functioning as a policy instrument to achieve reconfiguration of school priorities. The impact of the publication of such measures involves complex interactions, including variable interpretation by schools and teachers. All of this carries serious implications for students, teachers, schools, educational leaders and policy makers. A major review of performance indicators in 2010-12 saw a fundamental shift from ‘threshold’ measures to more complex, mixed measures. National qualifications (14-19) and national tests (5-11) are fundamental to the operation of these measures, providing the data to populate the national data sets upon which the measures are based. This has washback both into the education system and into specific qualifications upon which the accountability measures depend. Research into the use of accountability measures reveals positive and negative consequences. Current changes to the measures in England are designed to resolve some of the unintended consequences.

Coe and Heller-Sahlgren (Coe R & Heller Sahlgren G 2014 Incentives and ignorance in qualifications, assessment, and accountability in Tests worth teaching to. Heller Sahlgren G (ed) 2014 Centre for Market Reform of Education: London) recently have explored why system design is crucial for deliberate management of the impact of accountability. They list seven key criteria for accountability measures which relate to aspects of validity and fitness for purpose. Some research evidence supports the view that qualifications, assessment and accountability drive the curriculum and can incentivise schools and lead to improvement. However, it is important to understand and address the negative, unintended consequences that have resulted for a range of stakeholders.

Paul Petersen’s analysis of successive waves of reform in the USA highlights (Peterson P Friedman Lecture, (CMRE) Westminster 26 January 2016) the abandonment of outcomes-driven accountability as the principal federal reform paradigm, whilst other jurisdictions either scale up their use of such mechanisms (Anderson JA 2005 Accountability in Education UNESCO; OECD

2013 Synergies for better learning: an international perspective on evaluation and assessment OECD) or begin to consider using them as part of concerted action on improvement.

Shifts in accountability strategy in England are of interest internationally, since emphasis on different policy instruments–National Curriculum, national assessment, national examinations (at age 16 and 18), school inspection and accountability – have undergone fundamental shifts in the last decade. The shifts in US policy outlined by Paul Peterson have been manifest in arrangements in England, albeit with the details playing out in a slightly different way.
The structural link between assessment and accountability has played out differently in England, most recently with a sharp and unexpected de-emphasis on National Curriculum and a dramatic increase in emphasis on assessment and examinations.

The challenge for policymakers is to make the most of the potential positive impact whilst recognising and combatting against the negative effects. The paper explores instances of pressure exerted on tests and examinations through their use in accountability arrangements. It also explores ‘focussing effects’, where specific measures give rise to pressures on equity by encouraging teachers and schools to target specific groups as a means of securing improvement against measures, and where specific curriculum choices are constrained or encouraged by the measures. The paper acknowledges that accountability objectives such as measuring pupil progress is an important part of contemporary management approaches to educational systems. We argue that understanding the match between intention and effect, and mapping washback effects are vital for beneficial utilization of this type of policy instrument.

71. 'If you join them, you don’t have to beat them'

Birgitte Arctander Stub, Mari Bjugstad Wiken and Ida Large (Norwegian Directorate for Education and Training, Norway)

The joys and dilemmas of soft governance and best practice – reflexions from a policy level.

We will explore how Norway introduces and operates governance, based on soft and semi-soft management, low accountability pressure and few consequences, all mixed with a high level of institutional trust and interpersonal trust. What advantages and dilemmas do we experience? What are the challenges?

Norway has a population of around 5 million. There are some large cities, but mostly Norway had a rural structure of settlement. Schools are administered by 428 municipalities and 18 different counties, with a comprehensive mandate of local decision-making. The Norwegian Directorate of Education and Training provides governance, guidance and supervision, while the schools and quality of the schools are controlled locally. What are the challenges?

We have a school system that is, to some extent, used to being led, but not particularly controlled. Our system is mainly based on trust, high motivation, corporative mechanisms and soft means. We have a long tradition of seeking broad political consensus, and have procedures that allow political interest and pressure groups a place at the table. This mechanism is present both on a centralized and local level. In schools, both teachers and pupils are allowed to – and expected to – participate and to be heard in discussions.

Since the beginning of 2000, Norway has introduced several new components and development programs to enhance quality assessment and quality development in schools. Almost all of the new components are low stake, and have relatively soft incentives. The assessment regulations have a pedagogical tone. School inspections come with guidance and support, and several possibilities to righten things with supervision. National tests,with their heavy load of formative information for use in classrooms, and exams, are developed centrally, but assessed locally by teachers. Schooling and training the local external examiners helps to add an common assessment practice for teachers. If a school or municipality fails to improve quality-wise after having taken part in a development program, they will be offered the chance of participating in a new program, rather than being held accountable for previous failures.

There are many dilemmas. According to OECD Education working paper no 108, 'trusting may have productive consequences for an individual, yet may or may not be beneficial to his/her society'.

We will only mention a few dilemmas at this stage. Because of the decentralisation, some might say we know too little about what is going on. We are also forced to accept a great deal of difference between municipalities and schools. This again means there is a great deal of difference in quality, and in what is offered to pupils. Here we can add a concern about this leading to a certain grade of cementation, rather than equalization. How much time should we accept that it takes to improve quality? And who do we blame when nothing happens?

72. Improving students' future prospects or extending the reach of the accountability framework? Investigating the impact of the English Baccalaureate on the educational landscape

Emma Armitage (AQA, United Kingdom)

GCSEs are high-stakes assessments, the outcomes of which increasingly serve a dual purpose, as indicators of both student and school performance. As a result, conflicts have arisen between assessment that serves student interests and assessment for the purposes of school accountability. The EBacc provides an interesting example of this tension. It reports the percentage of students who enter and achieve A* to C grades in a specific suite of academic subjects: English, maths, science, history or geography and an ancient or modern foreign language. Although not yet a formal accountability measure–there is no pass rate schools must achieve–there are plans to require 90% of students in future cohorts enter the EBacc (DfE, 2015), which would extend the focus of performance tables from grades attained to subjects studied. Perhaps more importantly, setting entry but not attainment targets runs the risk of re-introducing so-called 'perverse incentives', whereby schools enter the majority of their students for the EBacc irrespective of their individual aspirations or aptitude – although this may be mitigated to some degree by the introduction of Progress 8.

Political rhetoric surrounding the EBacc's introduction revolved around the need to raise standards and reduce educational inequality (DfE, 2010). Education ministers maintain that studying this collection of rigorous academic subjects will keep students' options open for the future and allow them to compete with their international counterparts, many of whom study a compulsory academic curriculum until the age of 16 (DfE, 2010). School leaders' and other education professionals, on the other hand, argue against a 'one-size-fits-all' approach to GCSE provision, especially one that may only be suitable for students who are motivated to study these subjects and capable of achieving good grades in them (Royal Society for the encouragement of the Arts, Manufacture and Commerce, 2016; SSAT, 2015). For students whose interests lie outside the academic sphere or who are not academically inclined, they question the wisdom of coercing them into taking the EBacc, in which they might perform poorly, rather than encouraging them to aim high in other subjects (NASUWT, 2016; SSAT, 2015; Wilshaw, 2015).

Despite these differences of opinion, little empirical work has investigated the impact of the EBacc on students' subject choices and academic attainment at GCSE. Hence, the aim of the current study was to examine how EBacc uptake and attainment have changed, taking into account prior attainment and school type. To this end, the GCSE subject choices and grades of four student cohorts – 2010 to 2014 – were obtained from the Joint Council for Qualifications (JCQ) and analysed using a variety of statistical techniques. The findings will be discussed with reference to the interaction between policy and practice, in particular the potential ramifications of the proposal to enter 90% of students into the EBacc alongside the introduction of new headline performance measures: Progress and Attainment 8.

Session Y: Addressing equal opportunities regarding participation in assessment situations

73. Providing Cross-national Comparability of Test Results for International Assessment in Higher Education

Elena Kardanova, Irina Brun, Denis Federiakin (Higher School of Economics, Russia) and Prashant Loyalka (Stanford University, USA)

Introduction
Cross-national research faces many methodological challenges; the issues related to measurement are the most crucial among those challenges. Comparability of results requires equivalence of measurement in different language and national versions of the assessments. In the presentation we demonstrate types and degrees of differences that may exist in different language versions of international assessments.

Methodology
Current study used data, obtained from ISHEL project (International Study for Higher Education Learning). The main goal of the ISHEL project is to study the quality of engineering education across Russia and China, for this study we used mathematics and physics tests. The data consists of 3.600 first and third year students from 21 undergraduate engineering programs in China and Russia.

We provide evidence in support of reliability and cross-cultural comparability of the assessment instruments. First of all, we analysed content and construct validity using cross-national expert evaluations of content areas and test items for each subject. Based on the results of the expert evaluations and the small pilot, the mathematics and physics tests were prepared. Each test included 45 items presented in multiple-choice format. The grade 1 and grade 3 tests for each subject had approximately 20 common items, thus making it possible to equate the test scores from different grades and place the results on a common scale (Reference 1).
Then we conducted psychometric analysis to ensure that both tests meet standards for educational instruments and can be equated between two grades and across two countries. We present differential item functioning (DIF) analysis results to provide evidence for cross-national comparability of test results and to ascertain the possibility of creating a common scale between the two grades and across the two countries. To test for DIF across countries we used different approaches. Firstly, we used the separate calibration t-test, which is a traditional approach within the framework of Rasch measurement. Secondly, we used ETS approach for DIF classification based on the magnitude of Mantel-Haenszel statistic. Thirdly, we used multi-faceted modelling. Further, we used Logistic Regression, which is also commonly used for detecting DIF. Lastly, we used IRT-ANOVA approach, which tested the group differences in residual scores.

Several methods of DIF analysis were used because preliminary analysis showed the presence of differences in ability distributions between Russian and Chinese students. We conducted simulation study to see if the differences in average ability and variance increased Type I error rate for DIF detection methods.

Results
Based on the analysis we showed that several items exhibit DIF. The ETS and IRT-ANOVA approaches were recognized as the most reliable and appropriate methods for our analysis. Based on the results of DIF analysis, 13 items demonstrated DIF across countries.

While the reasons for the observed DIF are not of immediate concern for this study, to create comparable assessment instruments, we decided to keep these 13 items and consider them as unique items for each country and to use DIF-free items for establishing the link between the two countries. We used simultaneous calibration procedures for creating a common scale between countries, which gave us a basis for making international comparisons.

References

- 1 Kardanova E., Loyalka P., Chirikov I., Liu L., Li G., Wang H., Enchikova E., Shi H., Johnson N. Developing instruments to assess and compare the quality of engineering education: the case of China and Russia // Assessment & Evaluation in Higher Education. 2016

74. Patterns in the use of languages for differentiated learning of mathematics by primary school students in Luxembourg

Catalina Lomos (Luxembourg Institute of Socio-Economic Research (LISER), Luxembourg) and Amina Kafaï-Afif (Agency for Development of Quality in Schools, Ministry of Education, Luxembourg)

MathemaTIC is an adaptive digital learning environment which enables the differentiation of mathematics learning in terms of the languages offered to the student to practise mathematics. Since 2015, it is being piloted by over 100 volunteer teachers and 1700 students of grades 5 and 6 in 40 primary schools in Luxembourg. In this multilingual country, the three official languages Luxembourgish, German and French are also those actively used for instruction during the years of compulsory education. Luxembourgish is the language of socialization in early childhood education, German is the language of instruction in primary schools and French is taught as a subject from age 7. Considering that around 46% of the population is of foreign origin (Statec, 2016), with the largest foreign group being the Portuguese, followed by the French and other smaller groups (Italian, Belgian, German, Balkan/Ex-Yugoslavian, English, other EU and other non-EU), the challenges of multilingual education are also reflected in the outcomes of the educational process from a very early age. Indeed, it should be noted that around two-thirds of the students enrolling in compulsory education at age four, do not speak any of these three official languages (MENJE, 2015) and by the age of 8, around 17% of them repeat one grade. In addition, over 25% of the children do not reach the minimum level in mathematics (EPSTAN, 2011- 2013) and 3 out of 4 grade repeaters do not speak German at home, a factor which significantly influences student success in Luxembourg. Learning therefore becomes a struggle for those students who do not master the languages sufficiently enough to understand the subject content.

To address this multilingual challenge, all interactive materials and tasks in the MathemaTIC digital environment are aligned to the national curriculum and offered in German, French, Portuguese and English and tailored to meet the individual mathematics learning needs of students in the last two grades of primary school. In this presentation, we will focus on the ability of MathemaTIC to promote differentiated learning by observing the average patterns in the use of the different languages by students to learn mathematics. All items are available in the four languages and are of three types: discovery (theoretical and descriptive), recognition and understanding as well as application and analysis. They exist in written text, audio and video formats. Students are thus able to choose the language of learning mathematics for any item or they can change the language even for a short sequence in the written, audio or video tasks within items.

As much as the preliminary data will allow us, we will illustrate the profile of students (in terms of their characteristics) who switch languages as well as the languages into which they normally change. We will also refer to the duration of the change, the format and type of items or tasks for which students change the language while practising mathematics in class or at home. At this stage, we will not associate language use with student performance in MathemaTIC.

Considering that the language of instruction is German, these patterns in the use of German and the other three languages for differentiated learning of mathematics will be relevant for policy and practice. It is hoped that such information related to the learning paths of students will inform teaching and learning to better support classroom learning and increase student success.

75. Assessing individual participation in collaborative group work

Ayesha Ahmed (University of Cambridge, United Kingdom) and Ruth Johnson (AQA, United Kingdom)

Collaborative problem-solving skills are critical for students to prepare for the world of work. They are also powerful tools for learning. It is important that students from all social backgrounds are given the opportunity to learn and make progress in this area, yet this is not currently happening in many classrooms in England.

Our aim is to research and develop methods for teachers to use for valid assessment of individual students' collaborative skills in the classroom. Enabling teachers to identify progress and target lessons towards areas of need can help to ensure that students are taught the skills necessary for effective group discussions and collaboration. These skills are traditionally taught in the elite independent school sector but should be taught in all schools. Current educational policies have side-lined these skills in the English National Curriculum, partly due to concerns that they are hard to assess.

We will report initial results of a study investigating methods for assessing 15 year-olds working collaboratively in triads to solve a computing problem. The students are in the first year of a two-year Computer Science course, a subject that lends itself to practical group projects involving discussions.

Our first step was to develop our understanding of the construct by reviewing the literature and related assessments and by consulting experts. For example, we have been informed by the OECD's (2013) core competencies for collaborative problem-solving, by Care et al's (2015) work on the social aspects of participation, perspective taking and social regulation and by work on oracy (e.g. Mercer, 2015).

The process of designing assessment tasks has been led by the teachers, with guidance from the researchers in order to ensure that the tasks require enough collaboration and discussion to allow informative assessment of individuals' participation in group work. The rating schemes are informed by theory but are also empirically grounded using pilot data from group task interactions.

We are collecting video data of the group discussions and log files of computer keystrokes during task completion. Teachers are making live ratings and comparative judgements of individual students' collaborative skills and of the products of the group work. Students are making self and peer assessments of their performances within the group.

Validation of the assessment methods and tasks is being carried out by interviewing teachers and students throughout the process, from the construct mapping to the interpretation and use of results. We are also considering alignment with relevant aspects of the curriculum and related qualifications, and using comparative judgement to help validate rating schemes. To address reliability concerns, assessments are carried out by multiple raters.

Detailed analysis of the videos of group discussions using established coding schemes allows us to identify solution-critical interactions and the utterances leading to these. This in turn allows us to form conclusions about the interaction between discussion skills and the quality of solutions, informing our understanding of the construct and how it should be assessed.

Our research will give us a better understanding of how the assessment of the process of collaboration interacts with the assessment of the products of collaboration. We hope the resulting assessment materials will be used formatively by teachers and will encourage the teaching of these critical skills to all young people.

References

- Care, E. Griffin, P. Scoular, C. Awwal, N. & Zoanetti, N. (2015) Collaborative problem solving tasks. In Griffin, P. & Care, E. (Eds) (2015) Assessment and teaching of 21st Century Skills: Methods and approach. Springer.
- Mercer, N. (2015) Why oracy must be in the curriculum and group work in the classroom. Forum for promoting 3-19 comprehensive education.57(1)
- OECD PISA (2013) Draft Collaborative Problem Solving Framework
- http://www.oecd.org/pisa/pisaproducts/DraftPISA2015CollaborativeProblemSolvingFramework.pdf

Session Z: Comparative judgment

76. A methodology for applying students' interactive task performance scores from a multimedia-based performance assessment in a Bayesian Network

Sebastiaan de Klerk, (eX:plain / University of Twente, The Netherlands), Bernard Veldkamp (University of Twente, The Netherlands) and Theo Eggen (Cito/ University of Twente, The Netherlands)

Computer-based simulations are increasingly being used in educational assessment. In most cases, the simulation-based assessment (SBA) is used for formative assessment, which can be defined as assessment for learning, but as research on the topic continues to grow, possibilities for summative assessment, which can be defined as assessment of learning, are also emerging. The current study contributes to research on the latter category of assessment. In this article, we present a specific type of SBA, namely, a Multimedia-based Performance Assessment (MBPA). In an MBPA, students perform tasks in a realistically simulated virtual environment, that resembles their actual work environment. In our case, we have built an MBPA to assess the knowledge, skills, and abilities (KSAs) of confined space guard (CSG) students. A CSG supervises operations that are carried out in a confined space (e.g., a tank or silo). In the virtual environment, the CSG students can navigate through interactive images and on-screen icons, as they would navigate through a physical work setting. The goal for the students is to identify all tasks in the virtual environment, and then to correctly perform these tasks. In this way, students are immersed in the virtual environment and experience the feeling of performing their vocation as they would in real life. Of course, the interactive and complex behavior of students performing tasks in the virtual environment produces a lot of data or scores. The main question answered in this article is: How can students' interactive task performance scores, produced by performing in the CSG MBPA, be applied in a psychometric model for making valid and reliable inferences regarding students' KSAs? We therefore present a methodology for scoring the interactive and complex behavior of students in the MBPA. We address two specific challenges in this article: the evidence identification challenge (i.e., scoring interactive task performance), and the evidence accumulation challenge (i.e., accumulating scores in a psychometric model). Using expert ratings on the essence and difficulty of tasks in the MBPA, we answer the first challenge by demonstrating that interactive task performance in MBPA can be scored. Furthermore, we answer the second challenge by recoding the expert ratings in conditional probability tables that can be used in a Bayesian Network (a psychometric model for reasoning under uncertainty and complexity). Finally, we validate and illustrate the presented methodology through the analysis of the response data of 57 confined space guard students who performed in the MBPA.

77. D-optimal adaptive comparative judgement

Yaw Bimpeh (AQA, United Kingdom)

The method of paired comparison is perhaps the simplest way of presenting objects for comparative judgement. With this method, objects are presented in pairs to judges; for each pair, the judge has to decide which of the two objects they prefer. A drawback of the paired comparisons method is that the number of comparisons increases exponentially with the number of objects or scripts. For example, a selection of 20 scripts requires 190 pairs to be judged, while a selection of 50 scripts requires 1225 pairs to be judged. This places a limitation on the usability of the method for a large number of objects. To address this problem, several authors (e.g. David, 1988) have proposed other design methods that aim to reduce the number of comparisons.

Comparative judgement methods have been applied to a variety of educational purposes in recent years. These include examination comparability studies (Bramley, 2007), peer assessment of undergraduate mathematics (Jones & Alcock, 2013), teacher assessment of creative writing (Heldsinger & Humphry, 2010), and practical science (Davies, Collier & Howe, 2012). Pollitt (2004) outlined how comparative judgement could be used as an alternative to marking.

Adaptive pairing algorithms used for comparative judgement in educational assessments are often ad hoc with little or no formal basis. In order to reinforce fairness in educational assessment, it is important that there is a sound basis for any adaptive pairing of scripts.
The two most commonly used models for analysing comparative judgement data are Thurstone's model and the Bradley-Terry model. In practice, these models yield nearly identical scale estimates for complete data. However, for comparative judgement where not every comparison is made – leading to an incomplete matrix of preference counts – the Bradley-Terry model, unlike Thurstone’s model, applies directly to incomplete data under mild restrictions.

Optimal design theory is widely used in improving the design of tests in education. For example, in calibration, sampling designs and test designs. Optimal sampling designs have been developed for efficient item parameter estimation (Berger, 1994; Jones & Jin, 1994; Buyske, 1998; Berger, et al., 2000; Lima Passos & Berger, 2004), and optimal test designs have been studied for efficient latent trait estimation (Berger & Mathijssen, 1997; van der Linden, 1998). Optimal design issues have also been applied to computer adaptive testing (van der Linden & Glas, 2000). In the current research, the optimal design ideas are applied with the aim of reducing the number of comparisons needed for comparative judgement, while ensuring that there are no sets of objects that are not compared. This does not necessarily mean that all pairs of objects are compared directly.

There has been considerable research into the application of optimal design to paired comparison experiments (see, for example, van Berkum, 1987; El-Helbawy et al., 1994; Graßhoff et al., 2003, 2004; Kessels et al., 2006; Street and Burgess, 2004, 2007; Großmann et al., 2009). It should be noted that some of these publications deal with choice experiments in general (where judges evaluate sets of two or more objects). All of the papers cited here assume that the quality parameters in the Bradley-Terry model are known or assumed to be zero in order to derive optimal design – an assumption that is highly unrealistic in practice.

In this research, we analyse comparative judgement under the Bradley-Terry model and proposed D-optimal method for pairing. The optimal pairing is determined using the cyclic partially balanced incomplete block and general equivalence theorem.

78. Comparative Judgement and Scale Separation reliability: Yes, but what does it mean?

San Verhavert, Sven De Maeyer, Vincent Donche and Liesje Coertjens (University of Antwerp, Belgium)

Advances in technology and psychometrics direct more attention to the development of new assessment methods, losing sight of possible consequences of these methods. This may cause overconfidence in the reliability of measuring complex competences. We think that this is also a threat in Comparative Judgement (CJ). With this paper we will attempt to turn this overconfidence into realistic confidence by providing a more solid basis for the reliability of CJ.
In CJ the reliability is measured with the Scale Separation Reliability (SSR) from Rasch analysis (index of person separation, Andrich, 1982). It expresses the internal consistency of the scale (Andrich, 1982). In our opinion this does not say much about what the SSR actually means.
The interpretations that Bramley (2015) provided from reliability theory, that the SSR expresses the correlation between the estimated rank-order and the true rank-order, is easier to interpret. Bramley provided inter-rater and test-retest correlation as practical applications. It is remarkable that the few studies reporting inter-rater correlation (Jones, Inglis, Gilmore, & Hodgen, 2013; Jones, Swan, & Pollitt, 2015) did not link it to SSR. Therefore, a first goal of this study is to look into the meaning of the SSR by comparing it with a form of inter-rater correlations. We did this using data collected with the D-PAC tool. For several assessments we calculated the reliability. Per assessment we then randomly split the data in two subgroups based on assessors, estimated the ranking and reliability per subgroup and calculated a Kendall's Tau correlation matrix between the rankings of the whole group and both subgroups. These correlations were compared with the respective reliabilities.
The results appear to confirm the interpretations of SSR provided by Bramley (2015). This provides an initial basis for the meaning of the SSR as inter-rater reliability/correlation. However, the results are not unambiguous. Further scrutiny of the specifics of the respective assessment might provide additional insights.

A second goal of this study is related to adaptive algorithms in CJ. Namely, it is clear that adaptive algorithms can only be evaluated on their efficiency in the light of their reliability. But besides the theoretical notion of n(n-1)/2 and a rule of thumb of 10 comparisons per representation there is no literature on how much comparisons regular CJ needs before an acceptable level of reliability is reached. This reveals the need to clarify the change in reliability in regular CJ. We will therefore look into the evolution of the reliability during multiple assessments using CJ.

We used the same data as for the first goal. Per assessment the judgement rounds were calculated based on the number of comparisons per representation. The reliability and inter-rater correlation was calculated up until a given round to investigate its evolution with increasing number of comparisons per representation.
Also here the results appear to confirm the rule of thumb of 10 comparisons. But again further scrutiny will be needed to explain the differences in the results between assessments.

References

- Andrich, D. (1982). Index of Person Separation in Latent Trait Theory, the Traditional KR-20 Index, and the Guttman Scale Response Pattern. Education Research and Perspectives, 9(1), 95–104.
- Bramley, T. (2015). Investigating the reliability of adaptive comparative judgement (Cambridge assessment research report).
- Jones, I., Inglis, M., Gilmore, C., & Hodgen, J. (2013). Measuring conceptual understanding: The case of fractions. Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education (PME 37) (Vol. 3, pp. 113–120). Kiel, Germany: IGPME.
- Jones, I., Swan, M., & Pollitt, A. (2015). Assing mathematical problem solving unsing comparative judgement. International Journal of Science and Mathematics Education, 13(1), 151–177.

Saturday 5th November

Session AA: Assessment accountability

79. Cross-validating teachers' judgments and test-based results

Carolyn Hutchinson (University of Glasgow, United Kingdom) and Sandra Johnson (Assessment Europe, France)

A recent OECD review of evaluation and assessment in education found increasing focus on evaluating students' achievements, often against curriculum 'standards' describing what pupils should know and be able to do at different stages in their learning. National tests and examinations are commonly used to determine the extent to which students are meeting such standards, and how they can be supported to improve outcomes. The same tests are also being used for system-level evaluation of teachers and schools, allowing comparisons amongst them and to hold policy makers and teachers accountable (OECD, 2013). O'Neill (2002) proposes 'intelligent accountability', which for Cowie and Croxford (2007) requires trust in professional judgments and focus on self-evaluation; and assessment and evaluation tools and processes that support student learning without distorting the purposes of education. Wiliam (2007) further proposed that there are three main functions for assessment: to support learning; to describe individuals; and to evaluate institutions and hold them to account. An effective system needs to separate the evaluative function from the other two.

Wiliam suggested that such a system might involve two stages. In the first, teachers' judgments about students' grades and/or levels would be based on evidence collected throughout a course, increasing validity by assessing a range of outcomes including constructs not easily represented in formal tests. For the second stage, teachers' judgments would be validated by national end-of-course assessments consisting of many short test units, written and practical, amongst them covering the range of learning outcomes for the course, with each student randomly allocated a small subset of the units to sample their learning. The test results would be used to validate and moderate rather than replace professional judgments about the levels or grades achieved.

In Scotland, a possible basis for such a system was already in place in 2007. The Government had proposed the use of moderated teachers' judgments for school evaluation, with the sample-based Scottish Survey of Achievement (SSA) providing information about achievement in different curriculum areas and core skills, at national and local authority level. Although the school-based assessment and the survey were complementary rather than integrated, nevertheless the parallels with Wiliam's proposals were striking. It was decided to further investigate the proposals with Year 9 pupils (14-year-olds), using test materials from the 2008 SSA mathematics survey.

Six high schools in six local authorities agreed to take part in the study, involving the each school's Y9 cohort. Teachers' judgments of their students' attainment levels in mathematics were gathered, before having the students take SSA tests. Testing took place simultaneously with the 2008 sample survey: several overlapping pairs of hour-long tests were administered at random among the students, each taking one test pair. As both teacher judgments and SSA tests produced attainment classifications for students on the same criterion-referenced levels scale, the resulting attainment distributions could be compared school by school.

The presentation overviews the results of the study, and concludes with a discussion about the potential of the approach to validation of teachers' judgments to address some of the issues and concerns surrounding the use of national and standardized measures for multiple purposes, including evaluation for accountability.

References

- Cowie, M., & Croxford, L. (2007). Intelligent accountability: ‘Sound-bite or sea change’. CES Briefing No. 43, June 2007.
- O’Neill, O. (2002). A Question of Trust. BBC Reith lectures, 2002. Accessed April 2016 from http://www.bbc.co.uk/radio4/reith2002/lectures.shtml
- Organisation for Economic Cooperation and Development (OECD) (2013). Synergies for Better Learning: An international perspective on evaluation and assessment, Paris: OECD
- Wiliam, D (2007) Designing an Assessment System: Presentation to the Scottish Qualifications Authority, August 2007. Accessed April 2016 at http://www.dylanwiliam.org/Dylan_Wiliams_website/Presentations.html

80. The transition in England from high stakes to low stakes assessment in primary school science

Oliver Stacey (NFER, United Kingdom)

Background

There is considerable debate about the relative benefits of sample testing as an alternative to full cohort testing in educational assessment (Childs and Jaciw, 2003). In England the science assessment at the end of primary school offers a case study into how school behaviour and pupil performance change as the nature of an assessment changes from a high stakes full cohort assessment to a lower stakes sample test.

Testing before 2009

Prior to 2009 the primary science curriculum in England was assessed as a full cohort test at the end of the final year of primary school (age 10-11). The test was high stakes with pupil results contributing to primary school accountability measures.

However there were a number of concerns among teachers and the science community about the detrimental effects this assessment was having on the teaching and learning of science. These concerns included a narrowing of the science curriculum in order to focus on preparing pupils for the test and concerns over the validity of assessing practical skills in a written test.

Partly in response to these concerns the full cohort test was abolished in 2009. In the intervening years the method of science assessment in the primary phase has undergone significant change, transitioning to a low stakes sample test used to monitor national standards.

Testing from 2010 onwards

From 2010-2012 a science test was administered annually to a sample of schools to monitor national standards. In these years all eligible pupils in a sample of 750 schools participated in the tests in order to monitor national standards in science at primary school.

Following the 2012 test it was decided to change the sampling methodology and assessment design to maximise the information collected from the assessments (Bew, 2011). As a result the assessment changed to a matrix sample design similar to that used in international surveys such as PISA. This design has a number of advantages over previous tests such as the ability to assess the whole curriculum in each test cycle. This change has helped to make the assessment a more valid measure of science attainment across the curriculum.

A number of other changes were introduced at the same time, these included:

- not giving schools participating in the sample pupil results
- selecting only 5 pupils per school as opposed to all eligible pupils within a school
- changing the timing of the test so that it happens later in the school term.
- giving schools less notice about being selected in the sample.

The effect on performance
There was only a slight decrease in the proportion of pupils reaching or exceeding the expected level in the test in 2010 after the test changed from a full cohort to a sample test. However there was a dramatic fall of over 20% in the proportion of pupils reaching or exceeding the expected level between 2012 and 2014 when the test changed to a matrix sample (DfE, 2016). It is likely that this large drop is due to several factors including reduced pupil motivation and schools devoting less time to science.

How this paper relates to the conference theme
This paper relates to the theme in a number of ways. Firstly it highlights how an assessment has been changed by Government in response to concerns from teachers and unions. Secondly it explores how an innovative assessment design has been used to address validity concerns with an assessment. Thirdly it illustrates the influence of international assessments on the design of a national assessment.

81. Policymakers Intentions for Test-Based Accountability Policy – the Test as an 'Omni-Instrument'

Lisa Amdur (Tel Aviv University, Israel) and Irit Mero-Jaffe (Beit Berl Academic College, Israel)

The purpose of the presentation is to describe the intentions of policymakers for test-based accountability policy. Test-based accountability relates to providers of education being responsible for promoting and ensuring student learning, learning which is evidenced in achievement measured on state-mandated tests. The context of the study was a state-mandated test known as the Meitzav that was introduced to the Israeli education system in 2002. The Meitzav included achievement tests in four core subjects – Hebrew/Arabic (L1), Mathematics, Science, and English (L2), together with a battery of instruments for evaluating the pedagogical setting and school climate. The Meitzav test was administered to fifth and eighth grades in all schools once every two years. Data for the present study was collected during its third year within a wider study that included school principals and teachers.

Qualitative data was collected from 24 policymakers through in-depth interviews and through the analysis of relevant documentation. The data was analyzed through the identification of codes that defined units of meanings. Codes with common links served to create categories. Categories were further iteratively refined to produce themes. Three key themes emerged – justification for the test, intentions for the test and presentation of the test to schools.

Findings indicated that policymakers felt a need to justify the introduction of the test and two types of justifications were identified. One type of justification was based on the historical context of state-mandated testing as the Meitzav was simply a continuation of national tests that had preceded it. Another type of justification was students' poor performance on international comparison tests and the country's resultant low ranking in relation to other countries.

Policymakers' intentions related both to school-based accountability and national accountability. School-based accountability policy intentions included: requiring schools to prepare action plans, requiring teachers implement the national curriculum, ensuring minimum standards for all students, and demanding that school personnel work harder and more effectively. National accountability, policy intentions included resource allocation, school evaluation and monitoring the status of education.

The findings indicated that policymakers also paid attention to how the policy was presented or communicated to the field. Two aspects of communication were evident, one relating to the explicitness of the policy message (ranging on a continuum from explicit to implicit) and the other related to the channel of communication (formal or informal). Policymakers conveyed both explicit and implicit content through and made use of both formal and informal channels of communication.

The findings are in keeping with what has been recognized in the literature which indicate that policymakers have multiple intentions for test-based accountability and these include school improvement (Fuhrman, 2001; Malen & Rice, 2008), awareness of instructional goals and expectations (McDonnell, 1994a), enhanced curricular planning (Hamilton & Stecher, 2006), holding schools and educators accountable for student achievement, (McDonnell, 1994a; Cohen, et al., 2007) and evaluation of school staff (Fuhrman, 2001; Malen & Rice, 2008; McDonnell, 1994a; Hamilton & Stecher, 2006; Cohen, et.al. 2007; Mintrop & Trujillo, 2007).

This study sheds light on the thinking behind educational policy decision-making. Having multiple intentions and aims, the most effective way of realizing these was to introduce a national test. Thus it would seem that policymakers perceive a test to be multi-purpose – a single instrument through which policymakers can realize a variety of goals. In other words, the national test may be perceived as being an 'omni-instrument' as it is the instrument of choice in promoting educational policy.

Session BB: The moderation of teacher assessment

82. Improving moderation of teacher assessed work

Chris Wheadon (No More Marking Ltd, United Kingdom) and Daisy Christodoulou (Ark Schools, United Kingdom)

In England, trust in teacher assessment is at a low ebb. Recent qualification reforms have seen a move away from teacher assessed practical work in public assessments due to the perceived difficulties of providing robust moderation systems. As a result, there is a risk that assessment will become more narrowly focused on academic work that can be easily assessed in examination conditions (Harlen, 2007). We present results of a trial of an innovative distributed approach to the moderation of teacher assessment. The approach uses recent developments in Rasch estimation algorithms to anchor Comparative Judgement sessions together (Hunter, 2003). The anchoring process is similar to test equating in that it uses anchor items across judging sessions to adjust the calibration of scores.

Teachers in England are required to assess their pupils' writing as part of the statutory assessments for the end of primary education. Their assessments form part of the schools' published accountability data. A sample of schools is visited each year by a moderating authority who consider the robustness of a school's assessment procedures. Preparation for potential moderation visits is time consuming, while the results of the moderation lack validation. Given that schools may develop different procedures there are obvious threats to accuracy and consistency of standards (Lamprianou & Christie, 2009). Different interpretations of national levels may lead to different standards being applied. External moderation may fail to pick up these different interpretations as moderators may be influenced by the teachers' interpretations of standards.

In the trial we will present, 6 schools judged portfolios of their pupils' writing with their own teachers using Comparative Judgement. Following their internal judging session, each school submitted a sample of their portfolios to a central moderating authority, who then undertook a judging exercise across all samples, again using Comparative Judgement. Representatives from each school formed part of the moderating authority. On conclusion of the judging, standards were established against the rank order.

Once standards had been established centrally, the scores of each schools' separate judging sessions were anchored to the moderating sample using a Rasch based equating procedure. The moderation process therefore automatically adjusted the scores of each school's judging sessions, and cascaded the standards outwards. The results of the trial were evaluated in the light of the pupils' results from other assessments as well as the results from traditional moderation. Data was also collected on the relative amount of time and effort expended by schools on the new moderation method compared with the existing moderation method. We will present the findings of the study, along with recommendations for implementing the procedure on a massive scale.

References

- Harlen, W. (2007). Criteria for evaluating systems for student assessment. Studies in Educational Evaluation, 33(1), 15–28. http://doi.org/10.1016/j.stueduc.2007.01.003
- Hunter, D. R. (2003). MM algorithms for generalized Bradley-Terry models. The Annals of Statistics, 32(1), 384–406. http://doi.org/10.1214/aos/1079120141
- Lamprianou, I., & Christie, T. (2009). Why school based assessment is not a universal feature of high stakes assessment systems? Educational Assessment, Evaluation and Accountability, 21(4), 329–345.

83. Thinking outside the box: Using e-marking to moderate internally assessed coursework through dynamic sampling

Matthew Glanville (International Baccalaureate, United Kingdom) and Thomas Kelly (RM Results, United Kingdom)

Seven years after the International Baccalaureate became an early adopter for e-marking of high-stakes examinations, they are once again pioneering the use of e-marking to externally moderate internally assessed coursework through a process known as dynamic sampling. This process was previously done manually using paper files and was an operationally laborious task, so the IB approached RM Results to see how they could use e-marking to streamline this process.

The process begins with the school uploading the 'initial' requested sample of coursework files for each component of an assessment into a secure portal run by the IB. The IB use a bespoke set of sampling and moderation algorithms to select which coursework files are requested. These are then uploaded into RM Results' e-marking tool, RM Assessor, where they are marked by an external examiner. The externally assessed marks are compared with initial marks given by the teacher, and should these marks fall out of tolerance, another 'enhanced' sample is requested from the school or institution. After this sample is marked, the IB can decide whether to moderate all internally allocated marks. In very rare circumstances a 3rd, much larger 'additional' sample can be requested.

As multi-media and audio files are now enabled for marking within RM Assessor, dynamic sampling can be used for all types of coursework and e-coursework. The digital allocation of files means that markers can be recruited for this process from all over the globe, vastly widening the available pool of examiners available to moderate. Using e-marking for this process has also enabled the IB to ensure that all files requested for moderation are marked by the same examiner. If an external examiner is unable to continue marking for any reason, the entire sample can be quickly and easily sent to another examiner to begin the process again and remain consistent. This is a critical requirement for the IB's dynamic sampling model as it provides transparency and ensures schools and institutions have confidence in this process. A number of other features have been made available in RM Assessor to enable the IB and examiners to closely monitor and control the marking of samples by school and sample level.

The whole process provides a highly effective and efficient means of digitally reviewing and moderating internally assessed coursework in a transparent and controlled way without the logistical challenges of doing this manually. RM Results has supported the IB in this process by enabling secure integration of IB systems and processes with RM Assessor.

In this session, the IB will share the methodology behind the tried and tested process for dynamic sampling and will discuss the challenges they faced and the benefits they are enjoying from this innovative process. They will be joined by RM Results who will explain how the e-marking element supports and enhances this process, and reduces operational strain for the IB.

84. Evidence for the reliability of coursework

Tom Benton (Cambridge Assessment, United Kingdom)

The reliability of an assessment is defined as the extent to which candidates' results would remain stable if the entire assessment exercise was repeated. In the context of a written examination this might mean understanding how much difference it would make to individual candidates if a different test version, with different questions, had been used and if examination papers had been marked by different markers. For coursework we may be interested in how much difference it would make if an alternative task had been assigned to students and another teacher had evaluated their work.

Whilst numerous studies (e.g. Bramley and Dhawan, 2012) have evaluated the reliability of written examinations, relatively little has been done to quantify the reliability of internal teacher assessment within schools (see Johnson, 2012, for a review). This is unfortunate since several high-stakes qualifications in England, including GCSEs and A levels, depend upon the reliability of internal assessment, often in the form of teacher-marked coursework (or similar). The use of coursework is justified in terms of ensuring the validity and authenticity of the learning experience, however, quantitative evidence of reliability is often lacking. This paper will attempt to infer something about the overall reliability of coursework by comparing its value to that of written examinations taken at the same time in predicting future examination scores. Specifically the paper will explore the extent to which scores from coursework (or controlled assessment – a more tightly constrained form of coursework) and from externally marked examinations in GCSEs at age 16 in England are predictive of scores in both coursework and external examinations at A level two years later. The study focuses on scores in History and English Literature GCSE. For History, separate analyses were conducted for candidates taking their GCSEs in 2008, 2009 and 2010 and taking subsequent A level assessments in 2010, 2011 and 2012 respectively (approx. 3,500 candidates with matching GCSE and A level in each pair of years). For English Literature separate analyses were conducted for candidates taking their GCSEs in 2012 and 2013 and taking subsequent A level assessments in 2014 and 2015 respectively (approx. 1,500 candidates in each pair of years). For both pairs of years analysed for English Literature, scores on GCSE coursework were more predictive than GCSE examinations scores of achievement on both A level assessments (whether coursework or externally marked examinations). In addition in each pair of years analysed for History, GCSE coursework scores were more predictive of achievement in A level coursework than GCSE examination scores were. However, in two out of three cases for History, GCSE exam scores were more predictive of A level exam scores than GCSE coursework. An explanation for this result could be found by observing that students going on to study A level had a far more restricted range of coursework scores (compared to the full GCSE cohort) than is the case for GCSE examination scores; thus restricting the predictive power of the coursework scores. Once this restriction of range was accounted for it was again found that overall coursework was just as predictive as externally marked tests in forecasting future examination performance. Since reliability is a necessary pre-condition for (predictive) validity, these results suggest that coursework scores may be more reliable than is often recognised.

References

- Bramley, T., and. Dhawan, V. (2012) Estimates of reliability of qualifications. Chap. 7 in Ofqual's Reliability Compendium, edited by D. Opposs and Q. He, 217-320. Coventry: Ofqual/12/5117.
- Johnson, S. (2012) A focus on teacher assessment reliability in GCSE and GCE. Chap. 5 in Ofqual's Reliability Compendium, edited by D. Opposs and Q. He, 365-416. Coventry: Ofqual/12/5117.

Session CC: Assessment – social implications

85. Developing an assessment for 4 year-olds – challenges and tensions

Heather Bamforth and Catherine Kirkup (NFER, United Kingdom)

This paper discusses the challenges faced in designing an assessment for 4 year olds on entry to school (a baseline assessment) that is valid and reliable. We also will examine the social and political climate within which this assessment was launched.

Educational context

In England, this is a time of significant change in both primary and secondary schools. What is being assessed, the standard that is expected and the measurement scale have all been revised. There has been a shift from the previous national system of using only attainment for accountability purposes to a system where progress is also taken into account (DfE, 2011). The current Early Years curriculum (DfE, 2014) focuses on the development of the 'whole-child' from birth to 5 and pedagogy is firmly rooted in children learning through play. Early Years assessment practice is observation-based and the measure used nationally to 2015 was well-established.

Implementation of the baseline policy

As part of the changes in curriculum, standards and measures, in Spring 2014 the English government announced the introduction of a progress measure for accountability purposes from school entry at age 4 (Reception) to the end of primary schooling at age 11 (Year 6). Part of this would include the introduction of a baseline assessment to measure pupils' starting points on-entry to school. In order to give schools choice, the government proposed that several different providers would be selected and a range of suitable products made available to schools. The criteria the potential products had to meet included both content and technical standards. A total of six products were deemed to meet the criteria and they then needed to secure a 'market share' of at least 10 per cent. Each supplier offered a different style of assessment. Three products received enough orders to continue to the pilot 2015 and the government funded these products for use in autumn 2015. As part of the pilot, the government undertook a comparability study (DfE, 2016). The similarities and differences of the baseline assessments will be explored in terms of meeting the DfE criteria and also the results of the comparability study. The latter, in particular, had an impact upon government policy which will be explored.

Challenges

In order to create the assessment, the research team considered the following:

- Validity – current research informed the selection of areas of literacy, mathematics and aspects of personal, social and emotional development have the strongest indicators for later outcomes. We also scrutinised the existing 'Early Years Outcomes' and other early years assessments

- Reliability – as the assessment was to be used for accountability purposes, it was important that the assessment was reliable, one part of this was preventing a ceiling effect
- Manageability – the age of the children involved meant that the assessment had to be relatively short and specific; we also added discontinuation and routing rules to ensure that it was accessible to most children.

Impact

Unsurprisingly, some interest groups were strongly opposed to the notion of assessing young children, persistently referring to it as 'testing'. These groups were vocal and emotive in their messages.

Despite opposition in some quarters, feedback from early years teachers suggests that they found the assessment manageable and the outcomes useful. Adjustments have been made to the supporting documentation (in particular reporting outcomes) to further assist practitioners. The results of the comparability study has meant that the assessments cannot be used for accountability purposes and the implications of this will be considered.

86. Reconsidering young people as social and political agents in educational assessment reform: Students' voices in national assessment transformations

Jannette Elwood (Queen's University, United Kingdom)

Assessment systems are a major driver of educational change and across nations, they are continually modified more often for political than educational gain. As the theme for this conference highlights, the consequence of these modifications to national and international educational assessments show them to be more complex socio-political phenomena than neutral, technical evaluations of students learning. Assessment is thus shaped and transformed by social and political agents interacting at the local, national and international level. Such a position requires us to rethink who are the social and political agents interacting and mediating assessment policy and practice – this paper suggests that students are such agents and as a consequence they are both policy actors but also 'the acted upon' – that students are shaping educational assessment policy within their own contexts but are also being shaped by it through assessment opportunities made available to them and ultimately the outcomes they receive.

In this context, the paper takes as its focus the issue of young people's participation in national assessment reform as well as their views and perspectives on high-stakes examinations systems about which they are rarely consulted. Established and emerging data sets will be drawn on that have considered (and are considering) these issues with young people from different jurisdictions. These studies from countries within the UK and Ireland were interested in bringing to the forefront students' views on the most significant assessment phases of their education, specifically examinations at the end of compulsory schooling. Students are positioned as 'social policy actors', significant players in the mediation of national assessment systems not just (but also) subjects in their implementation and completion. Key themes are highlighted relating to young people's experiences of examinations and assessment at this stage of education, in 'high-stakes-for-them' situations and in different educational policy contexts which are operating within close proximity to each other but which are providing very different experiences for the young people involved. The impact on students of national assessment reform in situ, as well as a consideration of what is problematic about the qualifications they take is also discussed.

I have argued elsewhere that listening to students in this way, about complex, high-stakes examining issues has a limited history within the educational assessment literature. The lack of 'student voices' in educational assessment debates is problematic given that so many of the underpinning decisions about suitable assessment regimes and practice reforms as well as their fitness for purpose are assumed epistemological positions that consider particular roles that assessment should play in enhancing and improving the educational achievement of young people. For students' views to influence directly educational policy, there is an imperative for us to interact with students as social policy actors or agents in determining their own educational goals (Cook-Sather 2006). We know from last year's conference from a social justice perspective, that there will be differences in how students might experience assessment policy reform – these differences should be made visible to inform both national and international policies as well as institutional practices. The data presented in the paper highlights how young people feel positioned within assessment regimes and how considering their perspectives on these matters can improve our knowledge of our fundamental assessment practices and provide for us a fuller understanding of those elements that will promote high quality, fair assessment.

References

- Cook-Sather (2006) Sound, presence, and power: 'student voice' in educational research and reform. Curriculum Enquiry 34(4), 359-390.

87. Randomised Controlled Trials: how assessment research can contribute to evidence-based policy and social justice

Andrew Boyle (AlphaPlus Consultancy Ltd., United Kingdom)

Randomised controlled trials (RCTs) are an increasingly popular approach for carrying out educational research in the United Kingdom. Many researchers and policy makers contend that only by conducting well-designed experiments, including those using an RCT methodology, can policy truly benefit from educational innovations that have a sound evidence base. In particular, RCTs have been promoted by bodies that seek to promote the attainment of learners from disadvantaged circumstances.

However, our contention is that policy makers, general (i.e. not assessment) researchers and advocates of RCTs have thusfar paid insufficient attention to insights from assessment research. This is an omission; in particular, constructing a robust validity argument as is best practice in assessment research could enhance the validity of RCTs to provide meaningful information about the policy initiatives that the RCT investigates. This is because our discipline can give a framework for understanding the dependent (outcome) variable that can sometimes receive less attention in RCT planning and implementation than matters concerning the representativeness of samples of research participants.

The specific threats to validity in an RCT will of course vary from case to case, but in our experience, measurement concerns include:

- Construct validity: whether an assessed construct (for example oracy skills) can be measured sufficiently reliably to merit making fine-grained distinctions between control and treatment groups. Further, if certain 'esoteric' educational constructs cannot be measured sufficiently reliably, what does this say about RCT as a valid research method?
- Test method effect: would an assessment carried out using teacher assessment be a valid approach for an RCT? In particular, if the control group of teachers were required to be trained in assessing the construct, might this not 'bias' them insofar as being trained in how to assess a construct would also necessarily give them insight into how to teach the construct?
- Comparability of score meanings: designers of RCTs should be less sanguine in assuming that outcomes of educational assessments and qualifications are 'by definition' comparable. Where qualifications are provided by diverse providers (as is the case in the UK) one cannot assume comparability.
- Treatment of measurement error: inherent unreliability in test scores may be a problem for interpreting changes in RCT participants' abilities in the assessed domain. However, treatment of error components of score variance is far from straightforward; it may be legitimate (as some have contended) to effectively 'ignore' error variance (if it can be assumed to be equally prevalent on both wings of an RCT). But, equally, such an assumption may be flawed.
- Item calibration and scale length: when RCT participants sit an educational test before an

intervention, and then also sit a post-intervention test, the question of item and test calibration arises. In this area there are concerns around what happens if learners appear to not have made progress despite having engaged with the intervention, and the extent to which meaningful calibration can be effected across a wide range of abilities.

In raising measurement concerns such as those above, we will argue that assessment researchers can contribute to ensure the validity of RCTs, and thus, by extension, that our discipline can help to provide more defensible public policy making in education.

Session DD: Technical, political and social stakes when transferring to e-assessment

88. Utilising technology in the assessment of collaboration: A critique of PISA's collaborative problem solving tasks

Simon Child and Stuart Shaw (Cambridge Assessment, United Kingdom)

Technological tools are increasingly becoming embedded in learning, teaching and assessment. Advances in technology offer new opportunities for assessing collaborative learning and problem solving skills in areas and contexts where assessment would otherwise not be possible. Computer-mediated communication environments can provide a record of activity that can be kept, replayed, and modified.

This presentation divides into two parts. The first explores the extent to how technology can support facilitation and assessment of group collaborative learning, how technology can offer a more effective means for recording the capture of interactions between group participants, and how technology can be employed as an agent for initiating collaborative behaviour. The second offers an in-depth critique of how one institution (the OECD) through the Programme for International Student Assessment (PISA 2015) attempts to assess collaboration. In PISA 2015 students collaborate with computer-based conversational agents. These agents are designed to represent team members who exemplify a range of collaborative skills, knowledge and understanding, and are programmed to introduce a degree of conflict that needs to be negotiated by the human partner. Technology is used in an attempt to control interactional boundaries, with the intention of pinpointing collaborative behaviours and traits in students' responses. Furthermore, PISA uses technology in the recording of responses. In this critique, we present the outcomes of an exercise that mapped the assessment approach of PISA 2015 to pertinent constructs of the collaborative process, and recent theoretical developments related to engenderment of collaboration within assessment tasks (Child & Shaw, in press).
The fundamental constructs that comprise effective maintenance and progress of the collaborative state include: Social interdependence (when the outcome of individuals is affected by their own and others' actions); Conflict resolution (group members may take 'path of least resistance' when given a group task, either by avoiding conflict, reaching early agreement, or by dividing the workload into discrete sub-tasks where little discussion is required); Introduction of new ideas (creation of new ideas, speaking respectfully, keeping an open mind to new ideas, perseverance); Sharing of resources (sharing thinking and reasoning with others in the group); Cooperation/task division (group members sub-divide tasks without reference or discussion with other members); and Communication (using linguistic and non-linguistic features).

We also map PISA 2015 to five criteria that assessors should meet when devising a collaborative problem solving task. Some of these criteria relate specifically to the task itself, whilst others relate to aspects of group composition. The criteria include:

- Task is sufficiently complex – the common factor in all assessments of collaboration is that group members are set a problem.
- Task is ill-structured – a good collaborative task is one that cannot be solved by one capable

member of the group.

- Task should utilise technologies that facilitate the collaborative process – there are a number of ways in which technology can be introduced into a collaborative task: as a resource in information gathering; as a focus of the interaction; or as a collaborative partner.
- Group member dynamics engender negotiation – negotiation is unlikely if all group members agree on a solution to a problem, or if one group member forces their will or assumed knowledge onto another (e.g., in a tutoring scenario).
- Group is motivated to work together – in setting the task, the assessor needs to motivate group members to work together.

References

- Child, S & Shaw.S.D. (in press). Collaboration in the 21st century – implications for assessment. Research Matters 22.

89. Political stakes during the transition towards computer-based assessments: the case study of a large-scale online assessment of 160,000 students in France

Sandra Andreu and Thierry Rocher (DEPP, France)

Large-scale student assessment programs run nowadays in a transition phase, switching gradually from a paper-based to computer-based format. During this period, numerous issues are raised both from the theoretical and practical views, providing new opportunities while introducing new constraints. On the social and political level, these new assessment modes further beg the question of the use of the results. As computer-based assessments allow the evaluation of more students for a given cost, it is tempting to evaluate all the students all the time. As a result, the sample-based assessment model in France is liable to be called into question. A new policy model of standardized assessment needs to be conceived.

Historically in France, the standardized assessments are mainly used to account for the global results of the education system (Trosseille and Rocher, 2015). These large-scale assessment programs are thus more often carried out based on nationally representative samples and the results feed into national policies. However, the technology underlying computer-based assessments today enables the massification of these assessments without incurring any excessive additional costs. This opportunity responds to the ever-increasing demand from the local education authorities to undertake standardized assessments in order to obtain indicators for monitoring purposes.

Within this context, France set up a wide-scale computer-based assessment: 160,000 Grade 6 students took an online test in French and Mathematics. This assessment which is low-stakes for students is aimed at providing performance indicators for each of the 30 education authorities (academies). A representative sample was therefore selected from each of these 30 French authorities, corresponding to a high proportion of 1 out 5 Grade 6 students assessed.

In practical terms, this assessment required a highly-organized setup. A network of local IT platforms was created using the existing IT and statistic departments of the local authorities. Such a support was a key factor given the wide range of IT equipment which was incredibly striking, at least for France, even in secondary schools. The quality of internet access also varied widely hence undermining full-web test administration. However, in spite of these constraints and owing to the high commitment of the local actors, over 90% of students were finally able to participate in the assessment in good conditions.

As a result of the success of this large-scale operation, there is an inclination to extend the assessment to all the students. Indeed an obvious advantage is that of the operation costs which are potentially reduced by five as compared to a paper-based assessment, based on the experience of assessments undertaken in France. In addition, given the growing amount of equipment in schools, the implementation of such assessments may be foreseen as exhaustive for all students of one grade at a low marginal cost.

This is where the question of the use of these assessments come into play. The foundation of these assessments: the move from sampling to total population coverage inevitably leads to the need of reexamining the assessment policy. In the light of the known misuses related to this type of exhaustive assessment, a pilot test was undertaken in parallel to the wide-scale assessment in two education authorities, where all the students of 'priority education' carried out the assessment. Work is being undertaken with heads of the local education authorities to come up with a model of the use and publication of the results at the local level. The aim is to produce results to improve teaching practices and avoid the potential misuses of the assessment.

References

- Trosseille, B., & Rocher, T. (2015). Les évaluations standardisées des élèves. Perspective historique, Education et Formations, 85-86 , 15-35.

90. Learning in Digital Networks – A novel assessment of students' ICT literacy

Fazilat Siddiq (University of Oslo, Norway) and Perman Gochyyev (University of California, Berkeley, USA)

The permeation of information and communication technology (ICT) in society has forced changes in employment and education. Researchers have accentuated that there is a need for a broad range of skills that young people today need for navigating in education and workspace, and labeled them as '21st century skills' (Griffin, McGaw, & Care, 2012). Within this context, the transition to social media and social networks has been emphasized, and recent frameworks for assessing students' ICT learning with a particular focus on social networking and learning progression have been outlined (Wilson, Scalise, & Gochyyev, 2015). Moreover, research has pointed out the lack of innovative and authentic tests for measuring competences such as communication, collaboration and problem solving (Siddiq, Hatlevik, Olsen, Throndsen, & Scherer, 2016).

Against this background, this study evaluates a novel test, the Learning in Digital networks – ICT literacy (LDN-ICT) test, which aims to measure ICT literacy, with a special focus on students' interaction while solving problems in digital networks. Students are given a range of tasks related to different subject contents (e.g., mathematics and natural sciences, and social sciences), each of which they are asked to solve collaboratively through two different platforms (e.g., GoogleDocs and chat). The LDN-ICT test was initially developed by the Assessment and Teaching of 21st Century Skills (ATC21S) project and was translated, revised, and adapted to the Norwegian language and school culture. We specifically investigate the reliability and validity of the instrument by means of Item response theory (IRT) modeling framework. Data were collected from a sample of 175 Norwegian students in grade 9.

Our analysis revealed good weighted mean square fit statistic for the 39 items, using unidimensional IRT (Rasch model). However, given the multidimensionality of the underlying framework, we used multidimensional Rasch model to account for the dimensionality structure of the construct. As a result, we found that a four-dimensional Rasch model fits the data significantly better than the unidimensional Rasch model. Furthermore, differential item functioning with respect to gender and socio-economic status was examined, and no indication of the unfairness of the test was found. The four dimensions identified in the measurement model positively correlated with the constructs: ICT self-efficacy, learning strategies, and academic aspirations. These findings show the importance of students' self-beliefs but also point out the importance of possessing strategies to use ICT efficiently and for enhanced learning through digital networks.

Our findings provide strong evidence for the reliability and validity of the Norwegian version of the LDN-ICT literacy test with respect to its dimensionality, relations to other constructs, and the generalizability across subgroups. Moreover, the LDN-ICT literacy test serves as an example of an innovative test including authentic tasks in which real time student-student interaction is facilitated.

References

- Griffin, P., McGaw, B., & Care, E. (2012). Assessment and Teaching of 21st Century Skills. Melbourne, Australia: Springer. doi:10.1007/978-94-007-2324-5
- Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I., & Scherer, R. (2016, accepted). Taking a future perspective by learning from the past–A systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. Educational Research Review
- Wilson, M., Scalise, K., & Gochyyev, P. (2015). Rethinking ICT Literacy: From Computer Skills to Social Network Settings. Thinking Skills and Creativity. Doi:10.1016/j.tsc.2015.05.001

Session EE: Background factors influencing student achievement and the success of schooling

91. Academic achievement and subjective well-being in primary school children

Tatjana Kanonire (Institute of Education, National Research University, Higher School of Economics, Russia)

Students' academic achievement considered to be an important education quality criteria for all education participants – for students and parents, teachers, and policy makers. Depending on the assessment purposes it could be used for different decisions with low or high-stakes. Usually students' achievement is the only one criteria used for education quality evaluation on all levels. However, students' achievement is not the only one purpose of the modern education. In modern school students' well-being become very important. Academic achievement and well-being could be the result of combination of different personal and social factors.

Subjective well-being is considered to contain three components – positive affect and negative affect, and general life satisfaction (Bradburn, 1969; Andrews, Withey, 1976). But when the focus is on well-being in school context such components can be extended. In this study students' well-being is analysed as physical well-being, well-being at school (satisfaction with different aspects of school life and environment and general satisfaction), positive and negative affect toward school, and relationships with classmates.

Students' achievement is assessed by Student Achievement Monitoring (SAM) (Nezhnov, Kardanova, Vasilyeva, Ludlow, 2015). This instrument allows detecting different levels of academic mastery in Mathematics and Language based on Vygotsky's theory.

The purpose of the study is to analyse how academic achievement and well-being are interrelated in primary school children and what combination of personal (personality traits and academic motivation; controlling for demographical characteristics and intelligence) and social characteristics (parents' education and family income, school type) could better predict both achievement and well-being. The theoretically built model will be checked in the empirical study. To answer the research question data from students and their parents are gathered.
Student' achievement is measured by SAM. Well-being is measured by Brief Adolescents' Subjective Well-Being in School Scale (Tian, Wang, Huebner, 2015), Classmates' Friendship Relationships Questionnaire (Turilova-Mifičenko, Rafičevska, 2008) and additional items developed by author. Academic motivation is measured by Elementary School Motivation Scale (Guay, Chanal, Ratelle, Marsh, Larose, Boivin, 2010). Personality traits are measured by BFI-46-A

(John, Srivastava, 1999) and non-verbal intelligence is measured by Raven's progressive matrices. Parents filled the questionnaire about family income and their education. All data are gathered electronically. Data are in progress. The results will be analysed using regression analysis.

Implementation of both students' achievement and well-being in school monitoring and policy planning is discussed.

92. The progress of first-year school-children: looking for factors of educational inequalities in the beginning of schooling

Alina Ivanova, Inna Antipkina and Elena Kardanova (National Research University Higher School of Economics, Russia)

Research purpose
This study investigates the factors related to the educational progress of first-year school-children with particular attention paid to between-school effects and children's background components such as SES and parent involvement/investments. The goal of the research is to examine the role of between-school differences in the educational outcomes of primary school students.

Background of the research: The literature shows that school features may belong to the important predictors of students' outcomes along with cultural capital of families and their economic status (Reference 1). However, no studies of the school factors and their associations with students gains have been conducted on a representative Russian samples.

Instrument: The data for this study come from iPIPS (international Performance Indicators in Primary School) assessment. The instrument was initially developed in 1994 in the UK and adapted for usage in Russia in 2013-2014 (Reference 2). The assessment consists of two cycles with baseline assessment conducted in the beginning of the first school year (in this study, autumn 2014) and follow-up assessment conducted in the end of the first school year (in this study, spring 2015), so it allows to measure children's progress. The data selected for this study include the measures of children's basic cognitive skills in reading and mathematics, and extensive contextual information from parents and schools.

Sample
The large-scale assessment was conducted in Tatar Republic (which is a region in the central part of the Russian Federation) on a sample of about 1500 first graders from 37 schools. The sample was randomized and stratified according to the location of schools (city vs suburbs) and school type (regular secondary schools vs schools of higher status). Tatar Republic is a multicultural region with 53% of Tatar population and 40% of Russians.

Method
Hierarchical regression analysis is applied to assess the differences in children progress across schools in connection with family characteristics and school environmental factors. The two-level models were constructed in which pupils were nested within schools with a range of explanatory variables.

Results (work in progress): According to the results of the analysis, children starting level in math and reading, as well as family SES and pre-school training levels were among the most important predictors of the students' results in the end of their first school year. However, the analysis revealed the group of school with the lowest average socio-economic status and highest students' achievements. Two hypotheses were offered to explain the findings. The first one takes into consideration so called phenomenon of school resiliency, when students from a disadvantaged socio-economic background at schools being in challenging context demonstrate quite high level of achievement. Another hypothesis is regarding specific parental practices that may lead to better children's results despite the disadvantaged background of low family SES. Both ideas will be tested on a larger (5000 children) representative sample during the second wave of assessment which will have been conducted by the end of spring 2016. The new set of data will allow to repeat the analysis, clarify the findings and better explain the between-school variance.

References

- 1 Bondi Liz, Attainment at Primary Schools: An Analysis of Variations between Schools, British Educational Research Journal, Vol. 17, No. 3 (1991), pp. 203-217
- 2 Иванова А. Е., НисскаяА. К.Стартовая диагностика детей на входе в начальную школу и оценка их прогресса в течение первого года обучения // Школьные технологии. 2015. № 2. С. 161-168. (Russian: Ivanova A., Nisskaya A., The baseline diagnostics of first-graders and pupils' progress in their first school year. // School technologies. 2015, P. 161-168)

93. Development of reading skills in pre-school students: the role of parental investments

Marina Vasilyeva, (Boston College, USA), Alina Ivanova and Elena Kardanova (National Research University Higher School of Economics, Russia)

Research purpose: In the present study, we ask several research questions to take a deeper look at the relation between SES and children's cognitive development. Our Research question 1 is how the key features of SES – parental education and income – are related to children's school readiness in terms of their language and literacy skills. Our Research question 2 is whether the relation between parental education/income and children's language/literacy skills at school entry is mediated (a) by access to educational resources and (b) by engaging children in educationally relevant activities. Our Research question 3 concerns the role of parental beliefs in the development of children's skills. We plan to examine whether parental beliefs vary as a function of SES and, in turn, predict children's school readiness.

Background of the research: The relation between SES and children's cognitive development has been well established. The literature systematically shows that family resources and family cultural capital are strongly connected with students' academic skills at the start of school (Reference 1). This paper proposes to provide evidence on the link between parents' SES factors and children's school readiness by using a unique dataset from Russia.

Instrument: The data for this study come from the iPIPS (International Performance Indicators in Primary School) assessment. The instrument was initially developed in 1994 in the UK (Reference 2) and adapted for usage in Russia in 2013-2014. The data selected for this study include the measures of children's basic cognitive skills, including language and literacy, and extensive contextual information from parents and schools.

Sample: The target population for this study is children enrolled in the 1st grade of school in the capital of one of the Russian regions – Krasnoyarsk. The data was collected during the 2014/2015 academic year. The participants included 1438 first-grade pupils recruited from 27 schools (65 classrooms). The mean age of the sample is 7,38 years at the assessment's time in October 2014. The sample was approximately evenly divided by gender: 51% boys, 49% girls. The sample is representative for the city and stratified based on two parameters–school location (6 city districts) and school status (general regular schools or schools with higher academic status).

Method

Structural equation modelling is applied to assess the direct associations and mediated effects among parental education, income, home resources and activities, parental beliefs and children's reading skills at the time of school entry. Results (work in progress): According to our hypotheses we expect following results: (1) We expect that higher levels of parental education and income will be associated with higher scores on the language/literacy test administered to children at the start of 1st grade. (2) We expect that the effect of parental education will be stronger than the effect of income. (3) We expect that income effects will be primarily mediated by access to resources, whereas educational effects will be primarily mediated by the amount and diversity of educational activities in which children are engaged at home. (4) We hypothesize that parental beliefs will vary as a function of income and education and that they also will predict children's scores on the language/literacy test administered at the start of 1st grade.

References

- 1 Davis-Kean, P. (2005). The influence of parent educational and family income on child achievement: The indirect role of parental expectations and the home environment. Journal of Family Psychology, 19, 294-304.
- 2 Tymms, P. (1999). Baseline Assessment and Monitoring in Primary Schools: Achievements, Attitudes and Value-added Indicators. London, David Fulton Publishers.

Symposia | Abstracts

Thursday 3rd November

1. Error in high-stakes assessments

Chairs: Isabel Nisbet, Paul Newton, Beth Black, Sarah Hughes and Stuart Shaw (United Kingdom)

Actual or perceived errors in assessments which are high stakes for candidates or the wider public can provoke anger and demands for someone to be held to account. The proposed symposium starts with a conceptual paper ('Talking about assessment error') which examines the concept of 'error' in ordinary language and its links to related concepts such as 'mistake', 'blame', 'fault', negligence' and 'accountability'. It offers a taxonomy of types of error relevant to assessment. It then considers technical uses of 'error' , including the relation with '[in]validity' and links with notions of 'true score' and 'expert mark'. The two purposes of the paper are (a) to offer greater clarity and understanding of the terms used in public debate; and (b) propositions about what the public has a right to expect when things appear to go wrong with high-stakes assessments. It draws from examples in standardised tests in the USA, graded examinations in the UK and professional competence assessments in Europe (including assessment of the competence of doctors).

The second paper ('Tolerating difference of opinion') draws on experimental research by the exams regulator in England (Ofqual) prompted by data which suggested that even where marking is standardised using detailed mark schemes and sophisticated monitoring of examiners is deployed, this does not eliminate the potential for difference of opinion concerning the 'definitive mark' or 'true value' of any particular response. In England, if a school believes that a particular candidate may have been incorrectly graded, owing to inaccurate marking, it is permitted to request that the candidate's work be reviewed. The research looked at four different approaches to marking review and, led to a decision to alter the review process, to make it more resistant to the threat of treating tolerable difference as marking error. This paper will look at the extent to which difference of opinion concerning the 'true value' of a complex response can be considered legitimate and therefore tolerable, particularly in a climate of the increasing high-stakes of examination results for students, teachers and schools.

The third paper is entitled 'To review or remark: that is the question: detecting error when reviewing students' marks'. It describes a qualitative strand of research within one exam board seeking to examine how scripts are reviewed during the Enquiry process. One conclusion was that in most subjects reviewers tended to remark rather than review, despite instructions to the contrary. These findings raise questions about the difference between reviewing and remarking, the extent to which remarking should be made available, the relevance of 'tolerance limits' and the extent to which the process of review is detecting error in the original mark, as distinct from proposing a different one.

References

- Gardner (2013): John Gardner, 'The public understanding of error in educational assessment', Oxford Review of Education, 39:1, 72-92
- Ofqual (2015a). Research into alternative marking review processes for exams (Ofqual/15/5804). Coventry: Office of Qualifications and Examinations Regulation

1a Talking about assessment error

Isabel Nisbet (University of Cambridge, United Kingdom)

Talk of 'errors' in assessments which are high stakes for candidates or for the wider public is often heated. Alongside the technical use of 'error' by psychometricians is the language in which commentators react to incidents where something has gone wrong in the assessment process. That dialogue can damage confidence in public examinations and assessment in general. There is a pressing need for ground-clearing conceptual analysis to inform and structure discussion of assessment error in a socio-political context.

This paper will first consider the concept of 'error' in ordinary language, and its links to related concepts such as 'mistake', 'blame', 'fault', 'negligence' and 'accountability'. It will set out a taxonomy or types of error relevant to assessment, including: 'human error', (including variants associated with computer-marked standardised tests (eg 'programming error' and 'keying error'), 'system error' and 'latent error' (including actions or constraints which make it highly probable that human or system errors will occur). Examples of the different kinds of error will be taken from different stages and processes in standardised tests in the USA and graded examinations in the UK and professional competence assessments in Europe (including assessment of the competence of doctors). A distinction will be drawn between error in the assessment process and error in the conclusions drawn from the results or shortcomings in the suitability of an assessment for its intended purpose.

Turning to technical uses of 'error' in assessment contexts, the paper will consider the relation between 'error' and '[in]validity', the difference between 'marking error' and 'marking variability' (including the notion of 'tolerances' within which variations between the judgements of markers is deemed legitimate) and links to the notions of 'true score' and 'expert score'. There are also underlying philosophical issues about the relationship between 'error' and the paradigm of assessment as measurement, with the (sometimes questionable) ontological assumption that there is a 'thing' to measure that can be measured with greater or less accuracy.

The paper will also relate 'error' in assessment contexts to technical uses of the term in other contexts, eg 'measurement error', 'random error', 'inaccuracy' and 'imprecision' and the use of 'root cause analysis' after adverse incidents.

The purpose of this analysis is twofold: (a) greater clarity and understanding of the terms used can inform and structure public debate and reduce misunderstandings in exchanges between assessment experts, politicians and the wider public when things go wrong; (b) armed with greater clarity it is possible to offer propositions about what the public has a right to expect in high-stakes assessments and which kinds of assessment errors deserve the opprobrium they receive.

The paper will conclude that the discussion of error cannot be reserved to assessment technicians as a private discussion or withheld from wider public scrutiny for fear of damaging confidence in public exams and assessments. The public has a right to expect high-stakes assessments to be planned and managed in a way that minimises the risk of avoidable mistakes, honesty about the scope for legitimate variability in outcomes, objective investigations when a bad mistake is made and transparency about the conclusions and the action to be taken to avoid a recurrence.

References

- Department of Health (2000): An organisation with a memory: Report of an expert group on learning from adverse events in the NHS chaired by the Chief Medical Officer, HMSO, Norwich, 2000
- Gardner (2013): John Gardner, 'The public understanding of error in educational assessment', Oxford Review of Education, 39:1, 72-92
- Rhoades and Madeus (2003): Kathleen Rhoades & George Madeus, Errors in Standardised Tests: A Systemic Problem', National Board on Educational Testing and Public Policy, Boston, May 2003, downloaded at: http://www.bc.edu/research.nbetpp/statements/M1N4.pdf

1b Tolerating difference of opinion

Beth Black and Paul Newton (Ofqual, United Kingdom)

Examining boards in England use a combination of question types to assess the range of knowledge, skills and understanding covered by their qualifications; from selected-response format, with answers that are straightforwardly right or wrong, to long-answer constructed-response format, with answers that can only be marked by applying academic judgement.

Although criteria for judging complex responses are specified within mark schemes, and examiners are trained and standardised to apply those criteria consistently, this does not eliminate the potential for difference of opinion concerning the 'true value' of any particular response. This raises questions concerning the extent to which this potential for difference of opinion should be tolerated; and, if so, then how.

One way in which difference of opinion can be accommodated is through the marking review process. Although most GCSE and A level scripts in England are marked only once, schools are entitled to submit an 'enquiry' for any candidate whom they believe to have been incorrectly graded as a result of inaccurate marking. Senior examiners, who review these scripts, are required to 'correct' marks when they believe that the mark scheme has been applied incorrectly. However, when they encounter marks which differ from those which they would have awarded, but which are not necessarily incompatible with the mark scheme, they are supposed to let those marks stand, reflecting a tolerable difference of opinion.

The operation of the review process for GCSE and A level examinations has raised concerns (Ofqual, 2015a,b). Some stakeholders worry that review examiners treat marking inaccuracy as tolerable difference, denying candidates the higher grades which they deserve. Other stakeholders worry that review examiners treat tolerable difference as marking inaccuracy, awarding candidates higher grades which they do not deserve. In addition to the fact that schools submitted 121,000 more enquiries about GCSE and A level grades in 2015 than in 2014, a rise of 27 per cent, it is widely acknowledged that certain well-resourced schools are more likely to request reviews than others.

The possibility that these factors may have rendered the extant process not only flawed but biased led Ofqual to undertake a major programme of research. This compared four different approaches to marking review, which varied along a number of dimensions, including whether scripts were reviewed blind, how 'tolerable difference' was operationalised, and how many review examiners were involved. Evidence from this research, alongside evidence that review examiners were not actually applying the extant process consistently, led to a decision to alter the review process, to make it more resistant to the threat of treating tolerable difference as marking inaccuracy.

Given the increasingly high stakes associated with examination results (for both candidates and schools), the idea of tolerating difference of opinion concerning the 'true value' of any particular complex response is very challenging. During our presentation, we will discuss our research alongside deeper philosophical and socio-political questions, such as: Are examiners judging responses or candidates? Is it possible, in theory, to define a 'true value' for a complex, extended response (and, if not, then...)? Is it reasonable to expect even senior examiners to be able to distinguish between tolerable difference and inaccuracy? How can the idea of 'tolerating' difference be communicated effectively?

References

- Ofqual (2015a). Research into alternative marking review processes for exams (Ofqual/15/5804). Coventry: Office of Qualifications and Examinations Regulation.
- Ofqual (2015b). Consultation on: Improving Reviews and Appeals of GCSE, AS and A level

Marking; Withdrawing the GCSE, GCE, Principal Learning and Project Code of Practice; New Requirements on Setting GCSE, AS and A level Grade Boundaries. (Ofqual/15/5807). Coventry: Office of Qualifications and Examinations Regulation.

1c To Review or to Remark – that is the question: detecting error when reviewing students' marks

Sarah Hughes and Stuart Shaw (Cambridge Assessment, United Kingdom)

Definitions of error invariably relate to 'something done incorrectly' either through ignorance or inadvertence. Assessment error is 'any variation from the mark, score or grade that characterises the student's 'true' capability in the aspect of their performance that is being assessed' (Gardener, 2013, p. 73). An error in the marking of a student's examination script, for example, will result in an outcome which is not a fair reflection of the student's true level of attainment (Ofqual, 2013).

Once a student has received their examination results a school can request a check for any errors in the marking of their work. 'Enquiries about results' may be prompted by a genuine perception of error in marking. Enquiries may also be an attempt to raise outcomes even when there is no perception of error.

Schools can request a clerical re-check. A clerical re-check implies that either parts of the script were unmarked (suggesting negligence) or that the marks were totalled and recorded incorrectly (suggesting carelessness), or both. Schools can also request a review of marking which entails a check that the agreed mark scheme was applied correctly. A request for a review of marking may be prompted by a genuine perception of error in the marking process or an attempt to 'game the system' whereby the enquiries process is used in an attempt to improve outcomes. Between 2008 and 2013, for example, one international awarding body received ten times as many requests for reviews of marking than for clerical re-checks.

This presentation describes a qualitative strand of research which sought to establish how scripts are reviewed during the enquiries process. In particular, how Reviewers understand the enquiries process and their role within it, and the behaviours Reviewers exhibit during the process. The intention of the review process is to review the original mark rather than remark the script. There are two behaviours inherent in the reviewing process: the detection and correction of any error in the original marking; and the judgement of whether the mark scheme has been applied within an acceptable tolerance of marks.

Research showed that across all subject groups, with the exception of Languages, Reviewers tended to remark rather than review, despite instructions to review the existing mark. Most Reviewers appeared to be unable to distinguish between remarking and reviewing: remarking was thought of as integral to the reviewing process. This has implications for how error, where it exists, can be detected during the review process. The study findings raise a number of key concerns which this mini-symposium will seek to explore. For example:

- What is the difference between remarking and reviewing?
- Should reviewing entail remarking and if so how can this be operationalised for different subjects and question types?
- Is the application of tolerance relevant to both reviewing and remarking behaviours?
- Does remarking a script allow Reviewers to detect error in the original mark or does it merely afford an opportunity for the Reviewer to provide a new mark which is as equally erroneous as the original mark (Ofqual, 2015)?

References

- Gardener, J. 2013 The public understanding of error in educational assessment, Oxford Review of Education, 39:1, 72-92

- Ofqual (2015a). Research into alternative marking review processes for exams (Ofqual/15/5804). Coventry: Office of Qualifications and Examinations Regulation.
- Ofqual (2013) Summary: Public perceptions of reliability in exams. Summary of Chapter 17 in Reliability of Assessment Compendium, Ofqual.

Friday 4th November

2 **Social and political underpinnings of admissions procedures to higher education – the perspectives of five European countries *(Atrium B)***

Chairs: Giray Berberoğlu (Turkey), David Gabelaia (Georgia), Iasonas Lamprianou (Cyprus), Christina Wikstrom, Per-Erik Lyrén, Magnus Wikstrom (Sweden), Avi Allalouf and Naomi Gafni (NITE, Israel)

The following papers will be presented:

1. Major problems, policy implications and possible solutions in the university admission system in Turkey
2. University entrance examinations in Georgia
3. University entrance examinations in Cyprus: A battlefield for lobbies and political proxy wars?
4. Meritocracy vs. egalitarianism: validity challenges in selection to higher education in Sweden
5. The effect of politics on higher education admissions in Israel

Globally, there is a growing need to increase and diversify the number of students in higher education. The desire to widen participation is motivated by political, economic, institutional, and social justice concerns (OECD 2001, Archer, Hutchings & Ross, 2005). However, resources are limited and, therefore, not every candidate can be admitted to his/her preferred field of study. The degree of tension between these two conflicting forces seems to be related to the level of political involvement and the strength of public reaction towards the assessment measures used for admissions decisions. As the tension between the need to increase participation and the lack of resources rises, the perceived importance of the admissions measures grows, and public concern and political involvement increase.

In this symposium we present issues related to admissions to higher education from five European countries. In all five countries fairness, equity and increased participation of young people from under-represented groups seem to be the main concerns of politicians, the general public and professional psychometricians. In all five countries standardized multiple choice tests are used as one of the admissions tools to institutes of higher education.The presentations address the policy context within which debates around widening participation have occurred. Each of the countries has found its own solution and compromises. In each of them, the implementation of a specific solution has generated consequences which in turn require attention and modification of the existing system.

A main concern related to these consequences and raised by the authors of the various papers, is how the use of standardized multiple-choice selection tools for admission to higher education impacts on teaching in the school system. High-school teachers tend to focus on the multiple-choice item types and teach to the test, rather than enhancing understanding and imparting knowledge. It became apparent that the weight acquired by the University Entrance Examination brought with itself a certain responsibility not only towards the higher education system, but also towards the educational system at large. A related issue is that the tests have been given too large impact, and that many upper secondary students focus more on preparing for the test, than on their upper secondary grades, resulting in negative consequences for their learning.

Another issue is the demand of the public, especially parents, for transparency with respect to the tests – since the stakes are so high, parents expect to be fully informed about all the details pertaining to the admission process and the tests used. This results in the disclosure of numerous items to the public, which in turn leads to the need to develop more and more new items, thus increasing the assessment cost.

The unintended consequences of the admissions processes currently used in each of the five countries are driving a variety of changes in the content of the test, in the score calculation method and in the reporting of scores– changes that have been initiated by both politicians and psychometricians. The dynamics between the various forces at play in the field of higher education and the interactions between politicians, the public, test takers and psychometricians in each of the five countries will be discussed by their respective representatives.

2a Major Problems, Policy Implications and Possible Solutions in the University Admission System in Turkey

Giray Berberoğlu (Başkent University, Turkey)

The university admission in Turkey basically depends on academic success of students based on two-stage multiple choice entrance examinations conducted by the Measurement Selection and Placement Center (MSPC). The first stage is mainly used for screening purpose with the verbal and quantitative sections in the test. Students who pass the first stage take different combination of the five tests in the second stage based on their choice of higher education programs and abilities. If students can not pass the first stage, they have to wait one more year to take the test again. In this two stage system almost one third of the applicants are selected and placed in a higher education program. The bulk of the applicants and the low selection ratio make the college admission extremely competitive.

The political concerns about the admission system basically focus on test security and equity. Recently conducting the first stage of the examination twice in a year has become another important issue among the politicians in order to reduce the tension among the high school students. On the other hand, academics' concerns are more related to psychometric properties of the test scores as well as the format of the tests used, such as computerized adaptive testing and using open ended questions as a part of the assessment.

The secure test administration has priority among the other issues for the education policy makers, since millions of students are tested on the same day and same time throughout the country. On the other hand, equity for the college admission has become one of the major issues in the past years. Students with relatively better socioeconomic background can invest on education more than the low income families especially through private tutoring (Berberoğlu & Tansel, 2014). Even though the government banned private tutoring institutions it continues informally within the school system. Moreover, the competitive nature of the admission system based on multiple choice item format only has some unexpected reflections on the educational practices in the country. Teaching on multiple choice items became a common methodology in the school system. As a policy strategy, all the test items are released to public after the examination. This application seriously hinders the improvement of the psychometric properties of the test scores since each year a new test form is prepared and neither tests equating nor piloting new test items are possible (Berberoğlu, 1996). The computerized test administration is seen as an alternative approach for improving test security and psychometric properties of the test scores, but it requires quite a strong technical and theoretical background to implement (Kalender,2011).
In this presentation each of the issue mentioned above will be elaborated with the results of empirical research studies. The politicians are more concerned about the logistics of the admission system in Turkey. On the other hand, there is also a great concern about the quality of the tests being used and the negative impact of testing on teaching and learning practices at schools among parents, teachers and all the other stakeholders. The discussion will focus on providing possible solutions for education policy makers to improve the admission system in Turkey.

References

- Berberoğlu,G. (1996). The university entrance examinations in Turkey. Studies in Educational Evaluation. 22(4), 363-373
- Berberoglu,G.&Tansel,A. (2014). Does Private Tutoring Increase Academic Performance of Students?: Evidence from Turkey. International Review of Education.
- Kalender, (March 2011). Effects of Different Computerized Adaptive testing Strategies on Recovery of Ability. Unpublished Doctoral Dissertation. Middle East Technical University

2b University Entrance Examinations in Georgia

David Gabelaia (National Examinations and Assessment Center (NAEC), Georgia)

Centralized system of University Entrance Examinations (UEE) was introduced in Georgia in 2005. The main incentive for introducing UEE was the political will to eradicate corruption and ensure a fair playing ground. Since then, higher education admission, as well as the state funding has been based exclusively on UEE scores. Consequently, UEE had considerable influence on the educational climate in the past decade.

The UEE consists of several tests. Three of these – Verbal and Quantitative Reasoning, Georgian Language and Foreign Language – are compulsory. It is also compulsory to take a fourth subject test, chosen by universities based on their particular programmes of study. Universities may also assign different weights to the results of the four tests for the particular programme. The students choose their preferred programmes of study. When the results are in, the centralized algorithm determines who is admitted to which programme and who is getting what share of their tuition fees covered by the state funds.

Initially there was a great deal of anxiety in the society about the new system. However, after the first admission cycle was completed, the public trust and support of the system increased tremendously due to its success in addressing the plaguing problem of corruption and inequity that was previously thought unbeatable (World Bank, 2012). Since then, the National Examinations and Assessment Center (NAEC) responsible for preparing and administering the UEE has strived to live up to this high level of public trust and support, as well as to maintain the structural, psychometric and content integrity of the tests. This has proved a challenge at times, not least due to political pressure exerted by the governing bodies in attempts to adjust the system according to their political agenda.

One of the changes advocated by some of the universities with rather strong political lobby was to drastically reduce the number of tests taken for university admission. A radical version of this was to rely on the Verbal and Quantitative Reasoning test only and dismiss all the other tests. The idea was not unpopular in the public eye, since for the applicants it implied less preparation, less stress, and allowed for results to be available much quicker. On the other hand, such a change would have dramatically undermined the already deteriorated school system, which had quickly responded to the new exams by focusing on the multiple-choice item types and teaching to the test, rather than enhancing understanding and imparting knowledge. These arguments were put forward in public debates, as well as inside political lobbies and gradually it became apparent that the weight acquired by the UEE brought with itself a certain responsibility not only towards the higher education system, but also towards the educational system at large and particularly the fragile school system. Thus the attempts to abolish the subject exams were resisted, as well as further attempts that pressured NAEC to do away with open-ended items. The struggle is not quite over perhaps, since the problem that the government wanted to solve with these attempts still persists. From one viewpoint, the problem is in the increasing volume of private tutoring, which undermines the equal availability of higher education to the public; however, the same problem can be seen as the inability of the existing school system to adequately provide for the educational needs of society. Whether the political will is directed to long-term policy planning suggested by the latter view, or the short-term surgery with high risk factors afforded by the former view remains to be seen.

Reference

- The World Bank. (2012). Fighting Corruption in Public Services: Chronicling Georgia's Reforms. Washington, DC.

2c University entrance examinations in Cyprus: a battlefield for lobbies and political proxy wars?

Iasonas Lamprianou (University of Cyprus, Cyprus)

The system of public University entrance examinations (UEE) in Cyprus has always been an example of a social and political 'success story'. Socially, it has served the people well, because students from all strata of society can get access to free, public University education. Politically, the UEE system is frequently presented by politicians as an institution which has withstood the test of time and leaves no space for corruption due to its very thorough security procedures. Still, the UEE system is frequently the battlefield where powerful lobbies and opposing political parties attempt to solve their differences.

In the Cypriot educational system there is no deeply-rooted, large-scale testing culture (see Lamprianou, 2012). Still, studies have shown that students support the UEE system because of its non-discriminatory and objective nature (i.e., the same rules apply for all), although they often acknowledge that there are washback effects(Michaelides, 2014, p. 437). Similarly, Lamprianou (2009) describes that there is some mistrust by the public in Cyprus about certain details and technical aspects of the UEE, but as a whole, the examination enjoys a huge public support.

The UEE system has undergone great changes in the last 25 years. Until 1991, Cyprus did not have its own University, and all students were competing through UEE for a place at Greek Universities or for a place at one of the local colleges (some of them were very prestigious and offered guaranteed employment). In parallel, (or despite the existence of the UEE), other students were privately applying at Universities in other countries, usually taking the A Level exams or other similar exams. At 1991, the University of Cyprus was established and since then, the UEE has been functioning as a centralized system for the local University, as well as for the Greek Universities. By 2007, a second public University was established, the Cyprus University of Technology, and the UEE became a centralized system to handle entrance to the Greek Universities, as well to the two local public Universities.

Due to this highly centralized examination system, the local Universities have very limited (close to none) control over the criteria by which admissions are decided. As a result, there are some very important problems, such as:

1. candidates may be accepted to University Departments without being examined on some highly relevant subjects (e.g. a student may be accepted at the Department of Biology or at the Department of Chemistry without being examined at either Biology or Chemistry).
2. the candidates can choose examination subjects from a long list of subjects, albeit with some restrictions. The fact that students are often examined on very different subjects raises the issue of the non-comparability of their examination scores (see Lamprianou, 2009, 2012).
3. the public universities are not allowed to set their own examinations (in parallel to the UEE) in order to attract students from the private secondary education schools of the island. This reduces the pool of candidates and deprives the students of private schools – who are often very bright students – the opportunity to enter the public universities.
4. Access to private Universities is very loosely regulated compared to the access to public universities.

The role of politics and powerful lobbies will be discussed in order to explain the current status-quo of public University entrance examination in Cyprus.

References

- Lamprianou, I. (2012). Unintended consequences of forced policy-making in high stakes examinations: the case of the Republic of Cyprus, Assessment in Education: Principles, Policy & Practice, 19:1, 27-44, DOI: 10.1080/0969594X.2011.608348
- Lamprianou, I. (2009). Comparability of examination standards between subjects: An international perspective. Oxford Review of Education, 35, 205–226.

2d Meritocracy vs egalitarianism: validity challenges in the selection to higher education in Sweden

Christina Wikstrom, Per-Erik Lyrén and Magnus Wikstrom (Umeå University, Sweden)

In Sweden, there are two main routes to higher education, and also two selection instruments; the upper secondary grade point average (GPA) and an admissions test; the Swedish Scholastic Assessment Test, or the 'SweSAT'. The SweSAT consists of a number of subtests measuring verbal and quantitative knowledge, assumed to be important for academic performance. The main purpose of the SweSAT is meritocratic: to rank students according to their general knowledge in verbal and quantitative skills and, hence predict their academic performance. But the test has also an egalitarian ambition: to contribute to a broadened recruitment, by giving students who for some reason underperformed in high school a second chance to compete for attractive study positions. Therefore, the test is not used in combination with grades, but as an optional alternative for the applicants, giving them a separate route to higher education. The test is open for all, and can be taken unlimited number of times. The SweSAT is by many seen as a fair and an attractive option, and the number of test takers has increased considerably in recent years.

It is well known that several purposes are problematic to meet, which is the case with the test. An early concern was group differences, and especially gender differences, which was one of the reasons for making the verbal part of the test larger. However, in recent years, the focus has mainly been on the predictive validity of the test (Lyrén, 2009). Therefore, in 2011, the SweSAT was revised, with the purpose to improve its predictive strength by introducing sub-scores, in addition to the total score. The verbal and quantitative parts were made more balanced, with both sections scaled and normed separately. The intention was to use the new score information when deciding on a selection model relevant for specific programs. This possibility has yet not been made use of, so the effects have not been evaluated in full. The changes to the test have however made some differences in other respects. Information from an ongoing validation project shows that the psychometric properties of the test are good or acceptable, but that gender differences have increased while other group differences have remained, or decreased slightly. The predictive validity of the total score varies between study programs and groups of students.

The most recent discussion on policy level is now less concerned with how the test works and more on its role in the admissions system. It has been argued that the test has been given too large impact, and that many upper secondary students focus more on preparing for the test than on their upper secondary grades, resulting in negative consequences for their learning. A recent suggestion has been to introduce an age limit to the test, and to decrease the number of study positions allocated on the basis of the test. This will change who are being selected to higher education, since the test works differently for male and female test takers of different age groups, ethnicity and social background, which is discussed in this paper.

Reference

- Lyrén, P-E. (2009). A perfect score. Validity arguments for college admissions tests. Department of Educational Measurement. Umeå: Umeå University.

2e The Effect of Politics on Higher Education Admissions – Israel

Avi Allalouf and Naomi Gafni (NITE, Israel)

Over the past few decades, admissions to higher education in Israel has been based on two criteria, more or less equally weighted: (1) Bagrut (matriculation) scores (average of external tests administered by the Ministry of Education, and internal tests administered by each high school), and (2) score on the Psychometric Entrance Test (PET), comprising three domains: Verbal, Quantitative, and English. Over this period, members of the Israeli parliament and the Minister of Education have initiated changes to the Bagrut and the PET. Parliament members accumulate political capital among the public by being socially conscious, reducing gaps between sectors of the population and smoothing the way for students. Ministers of Education want to make a mark early on, knowing their tenure is limited. Most changes have been made to the Bagrut, for which the Ministry is directly responsible.

Major changes made to the Bagrut have affected the: (1) composition and content of the tests – focus areas of material that will be tested, which are announced several months in advance; a lottery, which reduced the number of compulsory tests, and offered students the option of writing a paper instead of the test; (2) process of calculating average scores – bonuses awarded for certain subjects, respective weights of external and internal tests; (3) test dates – transition from giving all tests in the final year of high school to spreading them out over three years, and from having one test date per subject per year to having two test dates so students can improve scores; (4) special accommodations for learning disabilities. These changes resulted in less stringent requirements which led to higher scores and a rise in the number of students earning a matriculation certificate.

The PET underwent fewer changes. The parliament continually censures the test's content and questions its necessity. Over the past 30 years, parliament members have introduced numerous proposals to abolish the test, but none were adopted. In 1983, the test comprised five multiple-choice subtests – Verbal, Quantitative, English, Figures, and General Knowledge. The test, seen as culturally biased, was widely criticized. In 1990, the Figures and General Knowledge subtests were eliminated, partly as a result of this criticism. Twenty years later, a parliament member garnered a great deal of support for a law to abolish the PET. She withdrew her proposal after NITE modified the PET to strengthen the connection between test scores and the proficiencies needed for academic studies. These changes included calculating three general scores (instead of one which is currently calculated), giving greater weight to either the Quantitative score (for acceptance to science studies and engineering) or the Verbal score (for acceptance to the humanities and law). Moreover, a writing task was added and several item types with lower face validity were removed from the Verbal and Quantitative subtests.

In parallel, Israeli higher education underwent changes. Rather than choosing from among six research universities, students have access to some fifty public and private institutions. This growth has led to more competition among the institutions, easing acceptance criteria and lowering the level of studies. This clear and troubling trend is not being addressed by the political system, which continues to approve new institutions and to reduce the standard of Israeli higher education.

References

- Allalouf, A. & Gafni N. (April, 2015) Israel: Revising the Psychometric Entrance Test – Adding an Essay-Writing Section and Other Issues. Paper presented at the meeting of the American Educational Research Association, in an invited symposium entitled: Revising Assessment Systems around the World, Chicago, IL.
- Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. Educational Measurement: Issues and Practice, 13, 12-21.

Discussion groups | Abstracts

Thursday 3rd November

1. **Educational assessment for the future: Asia-Pacific and Scandinavian contexts**

Tony Burner (University of South East Norway, Norway), Nhat Ho Thi and Duyen Tran (Hanoi National University of Education, Vietnam)

Globalization leading to a more linguistically and culturally diverse population, advancement in technology and new ways of knowing and doing, and development in research on how we learn and develop are all factors that require educational change. Educational assessment has undergone radical changes the last two decades. In several respects, the Asia-Pacific and the Scandinavian contexts are not that different when it comes to the recent developments within educational assessment. So-called lifelong skills are promoted in both regions. However, the proposed work with lifelong skills is not related to educational assessment.

We take Vietnam and Norway as the examples of the Asia-Pacific and Scandinavian contexts here. In Vietnam, there have been radical changes in educational assessment carried out in primary level. In 2014, the Ministry of Education and Training issued Circular No30 (MoET, 2014a), stipulating the shift from a summative assessment approach to a more formative assessment approach in primary schools. Assessment is carried out during teaching and learning, aiming to support student learning, there are no comparisons or ranking, no marking, and alternative assessments such as peer- and self-assessment are encouraged. Changes are also reflected in policy on foreign language assessment. More focus is placed on language ability, which is different from the past when the focus was more on language knowledge. In 2014, MoET promulgated the Vietnamese version of CEFR, a set of standards to describe Vietnamese foreign language ability (MoET, 2014b). The requirements for learners' foreign language ability at different educational levels are also imposed, resulting in noticeable changes in foreign language education at all levels. Furthermore, in 2015, the educational authorities mention eight lifelong skills to succeed in school and life in general. Those are independent learning skills, creative thinking and problem-solving skills, aesthetic skills, physical skills, communication skills, cooperation skills, numeracy skills, and ICT skills (MoET, 2015). However, nothing is said about the assessment of these skills. How can educational assessment promote and support these types of skills?

In Norway, the government issued new assessment regulations in 2009, enforcing self-assessment, useful feedback and student involvement by law from grade 1 to grade 13. An evaluation reveals that primary schools have reached furthest in developing their assessment practices compared to middle and secondary schools (Sandvik et al., 2012). A recent study in foreign language teaching in middle school shows that in-service courses and seminars on formative assessment inspire teachers and increase their research-based knowledge, but they fail in several respects to practice what they preach (Burner, 2015). Similar to the Asia-Pacific context, lifelong skills are promoted in Scandinavia. A government-initiated committee, The Ludvigsen Committee, has the last two years worked with defining a construct for students' future competencies, looking 20-30 years ahead (NOU, 2015: 8). The committee suggested four main competencies: subject-specific, competence to learn, competence to communicate, interact and participate, and competence to investigate and create. The committee also suggested social and emotional competencies to be included in all the four main competencies. However, similar to Vietnam, nothing is said about how educational assessment can promote and support these competencies.

We would like to discuss the factors that motivated changes in assessment policies, the role of legislation on educational assessment, with examples from the Asia-Pacific and Scandinavian contexts. In addition, we will discuss any new assessment practices that have grown from the legislative directions mentioned above, lessons learned and the way forward by relating the lifelong skills as proposed above to educational assessment.

2. National Improvement Framework and Assessment of Children’s Progress

Kit Wyeth, Jane Gallacher, Donna Bell (Scottish Government, United Kingdom) and Graeme Logan (Education Scotland, United Kingdom)

The purpose of the session is to explore how the development and implementation of the National Improvement Framework will improve Scottish education and close the attainment gap, delivering both excellence and equity. Historically, Scotland used criterion referenced assessments such as the 5 to 14 (Assessments) covering the main curricular areas. The 5 to 14 Assessments provided a consistent and structured approach to assessing progress which was used in almost all schools. There was a clear progression pathway with prescribed content at key stages. However, it restricted teachers’ creativity and flexibility and there were concerns that the delivery and design of the curriculum was being driven by the assessments. The introduction of Curriculum for Excellence was considered a landmark development when it was introduced in 2010 covering the age range of 3 to 18. The revised curriculum is distinctive in that it moves from central prescription towards more emphasis on professional judgement. It draws upon assessment guidance in Building the Curriculum 5 (Education Scotland, 2009) which sets out the principles of assessment, standards and expectations. A key expectation of Curriculum for Excellence is that it builds on the experience of assessment initiatives such as Assessment for Learning (Inside the Black Box, Dylan William and Paul Black, 1998) to better meet the needs of local communities. One of the strengths of Scottish education has been the use of national policy reviews. The Organisation for Economic Co-operation and Development (OECD) report: Improving Schools in Scotland (2015), sets a number of challenging recommendations including developing an integrated framework for assessment and evaluation: 'This framework has the potential to provide a robust evidence base in ways that enhance rather than detract from the breadth and depth of the Curriculum for Excellence. Given Scotland’s previous bold moves in constructing its assessment frameworks on the best available research evidence at the time. It now has the opportunity to lead the world in developing an integrated assessment and evaluation framework.'

There is also a need to strike a balance between the formative focus of assessment and the need for a more rigorous evidence base. The National Improvement Framework is designed to deliver on these recommendations and includes the development of national standardised assessments in aspects of literacy and numeracy at key stages. There are six areas or drivers of improvement, including:

- school leadership;
- teacher professionalism;
- assessment of children’s progress;
- school improvement;
- performance information.

The drivers of improvement will provide information on progress on national priorities including the poverty related attainment gap and improving the life chances for children, young people and families. The Framework was published as part of the Programme for Government and is being taken forward jointly by Education Scotland and the Scottish Government. The collaborative approach to educational developments is well established in Scottish education promoting new ways of working across classrooms, schools, local authorities and nationally to tackle educational inequity. The proposals for the Framework have been consulted on widely. The use of information graphics or 'dashboard' to share evidence of progress in a visual format is being developed. The seminar will provide opportunities for delegates to discuss with Graeme Logan, Strategic Director, Education Scotland and Donna Bell, Deputy Director from the Scottish Government Scotland’s development and use of the Framework to improve Scottish education and close the poverty related attainment gap.

Friday 4th November

3. Standard-setting/maintaining and public trust in national examinations around the world

Lena Gray (Centre for Education Research and Practice, AQA, United Kingdom), Tina Isaacs (UCL Institute of Education, United Kingdom), Jo-Anne Baird (University of Oxford, United Kingdom), Dennis Opposs (Ofqual, United Kingdom), Christina Wikstrom (Umeå University, Sweden) and Anton Beguin (Cito, The Netherlands)

Processes of standard setting and maintaining within curriculum-related assessments form a key strand of educational assessment policies and programmes, and debates about 'standards' are often at the heart of educational reform. Many countries use curriculum-related examinations to select learners for higher education, work and other study options. Some countries also use these examinations as tools to measure school system performance. The need for functional and easy-to-explain high-stakes assessments to allocate scarce resources influences how and on what students are assessed, and how those assessments are judged or graded. Consequently, standard setting and maintaining are key sites for public management of policy. In many countries around the world, standard setting and maintaining are subject to intense media pressure, and are sometimes the focus of public disappointment and distrust, leading to heated political debates about ‘dropping standards’ and ‘failing assessment systems’.

Given the high stakes nature of these examinations, it is surprising that the ways examination standards are conceptualised and operationalised differently across nations has not been given sufficient attention. This is an interesting area because globalisation has begun to impinge on assessment systems, but public examination standards are still largely a bastion of the local. The meaning of ‘standards’ differs between countries and the stated value positions and processes relating to assessment standards differ markedly. Use of external and school-based/ teacher assessment varies, as does the trust assigned to teacher judgement, whether in assessing or in setting standards. Moreover, how social justice is perceived as a component of assessment standards is dissimilar across countries, reflecting broader cultural differences. How policy and politics affect standards in different countries has not been articulated well. Standard setting policies and processes can be difficult to articulate and communicate within the assessment research community, let alone to politicians, parents and students, and the wider public.

This discussion group will provide information on the findings of phase one of a project that aims to examine critically policy positions and processes for assessment standards in a range of countries, drawing on analyses from in-country experts and researchers. Case studies will be presented in which presenters will briefly outline their country’s policy positions on standards, indicate the sorts of processes involved in setting and maintaining those standards, and highlight the key issues in public debates about standards in their country. The case studies will cover standard setting/maintaining in:

- England
- Sweden
- The Netherlands

Discussion with colleagues on the extent to which initial findings contrast with, or coincide with views of assessment standards in their context will help in moving the project forward.
The discussion will illuminate similarities and differences in conceptual bases, operational approaches, and outcomes for candidates in participants’ own contexts as well as in the cases presented. In foregrounding the social and political underpinnings of standard setting policies and processes, it is anticipated that the discussion will involve challenge to current theory on standards, as well as critical reflection on how national organisations approach standard-setting/maintaining.

The discussion group will be of interest to researchers, policy-makers and practitioners interested in assessment standards.

Reference

Baird, J. & Gray, L. (2016) The meaning of curriculum-related examination standards in Scotland and England: a home-international comparison. Oxford Review of Education, forthcoming.

4. Developing Teacher Assessment Capacity: Diverse Perspectives from around the World

Lisbeth Brevik (University of Oslo, Norway), Christopher DeLuca (Queen's University, Canada), Christine Harrison (King's College, United Kingdom), Carolyn Hutchinson, Kay Livingston (University of Glasgow, United Kingdom), Sandra Johnson (Assessment Europe, United Kingdom) and Claire Wyatt-Smith (Australian Catholic University, Australia)

Given the growing accountability context of educational systems throughout the world, there is a rapidly developing need to educate teachers in effectively using assessments to promote, monitor, and report on student learning (OECD, 2013). In many countries, classroom assessments are increasingly being used to measure pupil outcomes in relation to national curriculum standards with large-scale assessments being used to determine the systemic progression of students towards standards. Despite this accountability context, research indicates that teachers have limited assessment literacy and struggle to keep pace with changing assessment demands (DeLuca & Klinger, 2010; MacLellan, 2004; OECD, 2013). Accordingly, there have been repeated calls for supporting teachers' assessment capabilities focusing both on pre-service and in-service teacher education (Brookhart, 2011; OECD, 2011).

In this Discussion Group session, diverse perspectives on developing assessment literate teachers will be explored. Specifically, presenters from Australia, Canada, Norway, and Scotland will share their current research on assessment education, focusing on innovative and effective approaches to educating pre-service and in-service teachers. A summary of the research projects used to stimulate discussion in this session follow:

Research Project 1

Through a systematic review of historical and contemporary professional assessment literacy standards, this project establishes a theoretical basis for assessment literacy. Specifically, fifteen teacher assessment standards from six regions (US, Canada, UK, Australia, New Zealand, and Europe) were analyzed to understand shifts in assessment standards over time and across regions. Results identify eight themes for assessment literacy, which serve as a conceptual basis for teacher education and future assessment literacy research.

Research Project 2

This project provides an analysis of the integration of Assessment for Learning principles in a newly revised 5-year Master of Education programme in Norway. In developing the programme, the University attempted to design highly relevant assessment situations in which student teachers could develop their assessment literacy. In analyzing data from 143 student teachers, this presentation addresses key topics in European assessment education by examining the integration of assessment principles into teacher training.

Research Project 3

Since 2013, the Queensland College of Teachers in Australia has been leading a significant initiative to monitor standards in teacher education programmes. The initiative takes as its focus the formation of the beginning teacher as assessment capable over the course of their academic programme and professional experience episodes in Initial Teacher Education (ITE). Involving all higher education providers in Queensland, the primary aim of this research was to examine the nature and extent of opportunities in ITE programmes that developed teachers' assessment literacy, and the formation of their assessment identities.

Research Project 4
This project has worked with practising EU teachers for 4 years as they learn to implement formative assessment practices within the context of inquiry pedagogy in science curriculum. Critical learning from this project has been the required adaptation of formative assessment practices within inquiry-based pedagogies and the deliberate consideration for supporting teachers' implementation of formative assessment principles across diverse pedagogies and curriculum.

Research Project 5
This project examines teachers' developing assessment capacities in Scottish schools through career-long professional learning (CLPL). This research is premised on the understanding that developing teachers' assessment capacity involves understanding teachers as individual learners and the conditions that promote their learning, as well as knowledge about their students. This project therefore explores: (a) how teachers understand assessment in relation student learning, curriculum, and pedagogy; (b) how teachers' assessment capacity improves students' learning through CLPL; and (c) how teachers' CLPL can be sustained in schools.

Based on brief presentations related to these research projects, discussion in this session will centre on embedding assessment into teacher education programmes and strategies for supporting practicing teachers learning in assessment.

5. The quality of school-based vocational assessment

Andrew Boyle (AlphaPlus Consultancy Ltd., United Kingdom), James Morgan (Scottish Qualifications Authority, United Kingdom), Bernadette Dilger (Universität St. Gallen, Germany) and Jean-Pierre Jeantheau (Agence Nationale de Lutte Contre l'Illettrisme, France)

Following on from the successful vocational assessment symposium at Glasgow 2015, this proposed discussion group will provide a forum for participants to discuss the quality of school-based vocational assessment.
In particular, the focus of discussion will be on:

- school-based vocational assessment (rather than, for example, workplace-based assessment).
- the quality (validity and reliability) of the assessment, not about the qualification system or wider questions about national skills provision, or the economy, per se.
- assessment designs and design principles that enhance the validity and reliability of school based vocational assessment.
- whether quality in vocational assessment is 'the same' or 'different' from general assessment.

Four participants from England, Scotland, Germany and France will give short introductory talks, based on contemporary issues that affect their respective systems. These short expositions are likely to address matters such as:

- Reforms intended to integrate vocational and academic assessment systems.
- The extent to which different validity arguments can be made for vocational and academic assessments within the same assessment framework.
- How the context of school-based vocational assessment influences the validity of the assessment.
- The careful use of internal, teacher assessment and external examinations.
- Diversity of provision (include assessments provided by multiple regions within a country, or by multiple independent commercial bodies – awarding organisations).
- The nature of constructs assessed in vocational qualifications – such as competence – and such constructs' relationship with other assessment features, such as grading.

- The role of jurisdiction-wide formal assessments as a tool for standardising school-based assessments.
- Government trends to make vocational appear more similar to academic examinations, and the consequences of this for validity.

Attempts of vocational assessment advocates to make the case for the quality of such assessment, and the response that this has engendered from stakeholders.
We hope that this group will give rise to a thoroughgoing discussion around these complex topics. It is our aim to promote the discussion of vocational assessment as an end in its own right, and to integrate these discussions with the mainstream debates of assessment research at AEA-Europe.

Saturday 5th November

6. Are the influences from social and political agents beneficial and a necessity in the development and validation of educational assessments?

Anna Lind Pantzare and Pia Almarlind (Umeå University, Sweden)

In order to produce sound and valid tests with respect to the Standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) there are a lot of issues to consider and take into account. Especially when developing tests on a national level. In addition, there are often several stakeholders having an ambition to influence the tests in different directions. Sometimes these stakeholders agree, but often their requests are diametrically opposed and it is not unusual that the requests are not in line with a good measurement practice. In the midst are the test developing organisations, commissioned to develop products that are valid in relation to the aim/aims for the test and therefore being developed within a sound measurement practice but also accepted by all users.

As described in the theme for the conference this external influence from stakeholders on the tests is immense and sometimes described as only negative. Often there are politicians who uses educational assessments, like national tests or exams, to control the school system on the one hand but on the other hand using the tests to implement changes. At the same time the politicians are sensitive to reactions from the teachers, parents and other stakeholders since they are important groups of voters. In Sweden the debate is at the moment focused on the large number of national tests and the workload they entails to teachers but also students. In a recently published government-appointed inquiry (SOU 2016:25, 2016) it is suggested that the number of national tests should be reduced, that the remaining tests should be less extensive and that the tests should be easier to administer and mark, which probably will affect the validity of the tests.

These external influences could, from a test developing perspective, be seen as problematic since it often introduces (rapid) changes of the tests. On the other hand, one could argue that these external influences are necessary prerequisites to have an ongoing process in order to develop the tests so that they become even more cost effective, valid and seen as valuable for the users.

We think it would be interesting to discuss this complex system of, on the one hand, social and political agents trying to influence and change the national assessment systems and, on the other hand, the test developing organisations aiming to develop assessments that are valid. But at the same time these organisations are dependent of getting resources from the agents to fulfil the commission, which might affect which changes that are implemented and not.

This is a proposal for a discussion group based on the broad question posed as title. Below we have specified some themes that would be interesting to discuss getting perspectives from different countries and testing systems.

- How are the products, i.e. the tests, and the processes developing the tests affected by the influences from different social and political agents?
- Are there stakeholders having greater impact, and if there are, is it a necessity or a risk? Why? How?
- Finally, is it maybe necessary to have this continuous external validation of the tests in order to develop, strengthen and legitimise them or does it 'ruin the work'?

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.
- SOU 2016:25. (2016). Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning [A new national system for assessing knowledge.]. Stockholm: Utbildningsdepartementet.

Poster presentations | Abstracts

Friday 4th November

A. Lower secondary school teachers' and students' conceptions of assessment pusposes

Georgia Solomonidou (Independent, Cyprus) and Michalis Michaelides (University of Cyprus, Cyprus)

Introduction and value of research

A major change in approaches to assessment is the shift to view assessment not only as a means to an end, to determine measurement and thus certification, but also as a tool for learning. Assessment can only steer learning when there is a constructive alignment between learning, instruction and assessment. Assessment practices can be extremely powerful in hindering or enhancing students' learning. Current research into assessment as a tool to support student learning is increasingly focused on how this support is perceived. As much as policy makers want to implement some changes which are persuaded to be good for their educational system, if students and teachers do not perceive them as meaningful and helpful, probably they are not going to apply them. Changes in conceptions may even have to occur first before any practical changes occur in students' and teachers' practices because the way they perceive things, is closely related to how they will treat them. As conceptions on assessment, in Cyprus were largely ignored, this study was designed to examine and combine assessment conceptions by both teachers and students.

Methodology

Questionnaires for teachers (N=95) and students (N=599) in lower secondary schools were administered to collect responses on the competing purposes of assessment as operationalized in the Teachers' and the Students' Conceptions of Assessment Inventories (Brown, 2009, 2010). Based on an initial analysis the content and sample of the qualitative phase which included individual interviews with teachers (N=7) and group interviews with students (N=15) were determined. Thus an explanatory sequential design of mixed methods was used where qualitative data helps explain or build upon initial quantitative results.

Results – Discussion

The results of this study show that teachers and students appreciate assessment in all its approaches and see the different uses and purposes of the variety of assessments. Both seem to realize that assessment could have been useful for improving teaching and learning. They do not tend to see assessment as an irrelevant or a bad process nor do they commonly associate it with school accountability, a notion not very familiar in the Cypriot educational system. Mostly through the interviews, teachers seem to argue that they do not use any firm assessment strategies to alter gaps in learning and teaching primarily due to inadequate emphasis by the Ministry of Education and Culture, as well as inadequate training. They say that in practice it is totally up to each teacher to guide teaching after assessment occurs and ultimately to give guidance to students to further their learning. Then it is up to each child's willingness for taking action as not any firm strategies seem to exist. It is evident that teachers and students are willing to change their practices in order for assessment to be used for learning. Students especially seem to believe in challenges and that they cherish several approaches of assessment such as projects and assignments which make them feel like active participants. The same applies with teachers, while admitting to have positive feelings about assessment and that they demonstrate a willingness to shift their assessment practices to a more cognitive constructivist approach. Assessment conceptions appear to be aligned with current educational theories and the Cypriot educational emphases, but beyond stated intentions, practices are not always in agreement with conceptions, and evidence is still lacking on whether relevant policies have been enacted in this direction. Research that takes into consideration key stakeholders views may help design a more relevant policies which are more likely to be enacted by students and teachers.

B. Does the mode of standardisation matter? The effect on reliability of marking and marker perceptions

Lorna Stabler, Magda Werno, Sarah Hughes and Stuart Shaw (Cambridge International Examinations, United Kingdom)

Technological innovations in assessment are not limited to the administration of examinations, but also contribute to an increase in the use of computer-facilitated marking, training and management procedures in the context of large-scale educational assessment.
When incorporating these technological innovations in marking, it is crucial to ensure that procedures are associated with at least comparable or higher levels of marking reliability compared to more traditional approaches to ensure the validity of assessments.

Previous research has shown the potential benefits of online standardisation, including the flexibility of the process, opportunities for frequent and detailed feedback, availability of statistical methods for monitoring the quality of marking and identifying aberrant markers, and consistency of communication within the marking team (AlphaPlus, 2014; Chamberlain & Taylor, 2011; Tisi, Whitehouse, Maughan & Burdett, 2013). However, replacing face-to-face meetings with an online standardisation process can engender potential risks. Research has highlighted the negative effects of more impersonal remote communication methods, longer decision-times and impaired effectiveness of group interactions as a means of helping markers interpret and internalise the mark scheme (Alpha Plus 2014) and perceived disadvantages of online standardisation, such as the lack of rich social interactions, difficulties building and maintaining a community of practice and impaired communication within virtual marking teams as well as confidence in marking decisions and marker enjoyment (e.g. AlphaPlus, 2014; Chamberlain & Taylor, 2011). Nevertheless, the literature indicates that both modes of standardisation have comparable, statistically significant positive effects on marking accuracy compared to conditions where no standardisation took place.

This study used a two-fold methodology. Operationally available data from 2011–2014 June examination series was examined to establish whether marker reliability (measured as deviation from the mark awarded by the Principal Marker) varies with standardisation mode. In order to explore the subjective experiences and preferences of markers in relation to the two standardisation methods, the perceptions of twenty one markers involved in face-to-face and online standardisation in relation to one of the IGCSE (International General Certificate of Secondary Education) Geography components were analysed.

Analyses involved a within-subject design to measure the consistency of marking for a group of fifteen Assistant Markers (AEs) a specific IGCSE syllabus in June 2012 and June 2013. Additional analyses were also conducted on data from AEs who marked across the 2011 to 2014 June examination series to establish whether any patterns in marking consistency can be, at least partially, attributed to the mode of standardisation. In addition, data was collected from twenty one markers who completed a paper questionnaire which comprised items relating to their perceptions and experiences associated with face-to-face and online standardisation.

No sufficient evidence was found to conclude that marker reliability was directly influenced by the introduction of the online standardisation process in IGCSE Geography. However, there was a strong preference expressed by examiners for face to face standardisation. Standardisation mode was found to be instrumental to examiners' subjective perceptions and experiences of marking and the extent to which they felt valued, engaged, and confident in terms of internalising and applying the mark scheme.

A number of recommendations were made in relation to training provision, mark scheme development, and the use of more advanced technologies to enhance group discussions and improve the quality of communication and feedback to markers during the standardisation process.

References

- AlphaPlus Consultancy Ltd. (2014). Standardisation methods, mark schemes, and their impact on marking reliability. UK: Ofqual.
- Chamberlain, S. and Taylor, R. (2011). Online or faceğtoğface? An experimental study of marker training. British Journal of Educational Technology, 42(4), 665-675.
- Tisi, J., Whitehouse, G., Maughan, S. and Burdett, N. (2013). A review of literature on marking reliability research. UK: Ofqual.

C. The impact of lack of legislation on educational assessment: a case study

Joaquin Cruz (University of Jaen, Spain)

In Spain, where the boom in language testing is unprecedented, marketing arguments seem to be leading the choices of test takers in the first part of the 21st century. Some tests are held in massive venues such as hotels or trade fair parks which host thousands of candidates, rivaling major sporting events.

The washback from these tests has also been important in a country that relied heavily on traditional methods of language teaching, certification and accreditation. Catching up with the rest of Europe has necessitated profound changes in the mindsets of professionals and, even nowadays, at times, Spain seems to be stuck in second gear while the rest of Europe is working at full speed.

In this context, Spanish policy makers have passed seven different major educational laws over the past 33 years, each of them intended to substitute the previous one: LGE (1970), LOECE (1980), LODE (1985), LOGSE (1990), LOCE (2002), LOE (2006) and LOMCE (2013). To complicate matters further, there are 17 autonomous communities in Spain, all of which have their own laws on education (see section 4). The resulting variegation has hampered mutual recognition of foreign language levels. As a result, depending on the community chosen, the linguistic proficiency of two different students may vary by up to two CEFR levels in the same academic year. It is because of this lack of intra-regional standardization that it has become necessary to establish external language tests whose results can clearly be linked to the CEFR.

This over-legislation has led de facto to a situation in which the lack of continuity in the policies of the central government has opened the door for external private companies to certify language levels not only at Spanish universities but also in other high-stake procedures like state exams. These companies do not only have a background in educational assessment but also the expertise to assess huge numbers of candidates.

It is not strange to find that certificates issued by well-established private companies are recognized by the central government to certify language levels, while the certificates issued by public universities, which have proved to be high-quality, are not recognized.

This has given some private companies a tremendous power and it is clearly creating a washback effect in language teaching. Along with this, the perceived political moves towards privately-developed tests are defining, in a way that no other theoretical approach can, test takers' preferences for one particular brand in Spain. The industry of testing, as it is called in this papers, moves millions of euros worldwide and is now borrowing in Spain many practices typical of pharmaceutical marketing.

Cambridge, Trinity, ETS and, more recently the British Council are the main contenders in the Spanish industry of testing. In the case of Cambridge exams, for example, the expectation generated amongst platinum centers, distribution centers that conform to Cambridge's most ambitious business development programs, is just one measurement of the current industry of testing. Trinity tests, on the other hand, have gathered momentum in the Spanish market following their recognition by a number of important institutions. In Spain, Trinity tests have surpassed in popularity ETS and even the most popular suite of exams in the past twenty years in Spain, Cambridge. Leaving aside reliability and construct validity concerns. Once again, Trinity and its market penetration in Andalusia are just one example of the many tests which are nowadays ubiquitous around the world. With different degrees of penetration, we are now living the heyday of language testing in Spain and this is shaping Spanish legislation of educational assessment.

D. A Contribution of Case Study Tests to Predicting Performance on Real-Life English Language Tasks

Elena Sokolova, Elena Iureva and Oksana Shahhoud (Russian State Social University, Russia)

Introduction

Today we are witnessing the shift away from a grammar-focused tradition to real engagement with language in real-world context. The role of testing as a mean of stimulating students' cognitive activity, ensuring comparability of learning is increasing. This is because our understanding of what assessment is and how to assess effectively has also developed. Even the objectives of estimating have changed. It is not only simple evaluating a student's level. Assessment has a role in motivation and self-reflection. Assessment is known as a systematic process of making judgments, and consequently reporting results about the effectiveness (with respect to achievement of intended learning outcomes) of learning and educational processes or about individual students' progress toward attainment of established educational objectives (Laborda, J. G., Sampson, D. G., Hambleton, R. K., Guzman, E. Guest Editorial: Technology Supported Assessment in Formal and Informal Learning//Educational Technology & Society, 18 (2), 1–2).

Assessment has changed since, firstly, the content of modern language course has changed and, secondly, the nature of assessment has also transformed. It has happened due to the influence of the technology. It can provide new ways of assessing. It is possible to video students interacting in groups; get them to record audio files. They can develop their writing skills in blogs and wikis.

The given research considers a case study as one of competence-based assessment methods and tools, an innovative approach to evaluation of students' performance in teaching English for Specific Purposes (ESP).

Case Study as an Assessment Format

Students study ESP not only because of being interested in the English language but also because of having to perform a certain task in English, which will contribute to adapting to their work conditions in the future more easily.

Thus, nowadays, foreign language teachers use problem-solving instruction for developing students' cognitive skills and facilitating their engagement in meaningful and authentic learning experiences. Case study is a vehicle that has the potential to help toward effectively meeting learning and assessment needs. The role of cases in teaching Business English is not doubted. The case development project under consideration is aimed at applying cases in assessing graduates' language competence at final examination. This alternative to a traditional form of examination is absolutely new for the Russian higher education. It's being tested these days at Higher School of Economics, Russia at the State Final Examination of Bachelors in Management and includes the instruction for working with the case; the background, the standard procedures; the task itself, approximate variants of the best, standard and unsuccessful answers, the competence map of the case which refers competences under consideration to those prescribed by the state standard.

The appropriately designed case for testing the graduates' abilities is believed to facilitate assessing the students' knowledge within a certain practically- oriented context so that learners are able to reflect and adjust the acquired skills towards optimizing the real-life situation, solving the problem and applying their knowledge of professional foreign language.

Conclusion

So the given research is meant to encourage universities to design examination environments that go beyond traditional learning practice to leave the cards with questions behind and focus mainly on students' needs, interests and creativity and support the development of highly skilled professionals who feel free to solve problems, work on the projects and think critically and proposes that in order to prepare students for the future the very form of testing and curriculum must be redesigned. Besides, case studies give an excellent opportunity for the immediate feedback from the examiner who is involved in analyzing the situation at hand.

E. A review study to identify adaptive algorithms for increasing the efficiency of Comparative Judgement

San Verhavert, Vincent Donche, Sven De Maeyer and Liesje Coertjens (University of Antwerp, Belgium)

Comparative Judgement (CJ) has proven to be reliable in different domains (e.g. Heldsinger & Humphry, 2010; Jones, Swan, & Pollitt, 2015). However, already from the early years after it was proposed as an assessment method by Pollitt and Murray in 1995, it was observed that CJ needs a lot of comparisons to reach an acceptable level of reliability (e.g. Bramley, Bell, & Pollitt, 1998). It was thus reported that assessors find the method tedious and repetitive (Bramley et al., 1998). Therefore, Pollitt (2012a, b) put forward the necessity of for adaptive selection algorithms, that will increase the efficiency of CJ without affecting CJ's reliability.

Up to present however, a systematic research on adaptive algorithms to reduce the number of comparisons in CJ is lacking. Also in the domain of Computer Adaptive Testing (CAT), as in other relevant domains like Psychophysics and Paired Comparison literature, a comprehensive review appears to be missing. Therefore, we will present a first step in tackling this inefficiency. In a systematic review (Petticrew & Roberts, 2006) we identify adaptive algorithms potentially increasing the efficiency of CJ and this from a broad range of research domains. As such, this review study might not only provide information for CJ research. It might also (e.g.) further spur CAT research in developing new algorithms. Furthermore, it can help CAT practitioners in making an informed (evidence-based) choice in constructing tests.

In a first part, a taxonomy of the adaptiveness will be constructed. An exploratory review in the domain of CAT preliminarily revealed seven levels of distinction: (1) statistical paradigm (Frequentist versus Bayesian), (2) information measure based or not, (3) Type of information measure (Fisher Information versus Kulback-Leibler Information), (4) weighting or balancing, (5) flexible or static constraints, (6) inclusion of a random component and (7) stratification of not. All distinctions and their meaning will be further elaborated on in the paper. In a second part we will attempt to review the efficiency of these algorithms in their respective domain, keeping the CJ context in mind.

References

- Bramley, T., Bell, J. F., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone Paired Comparisons. Education Research and Perspectives, 25(2), 1–24.
- Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain

reliable teacher assessments. The Australian Educational Researcher, 37(2), 1–19. http://doi.org/10.1007/BF03216919

- Jones, I., Swan, M., & Pollitt, A. (2015). Assing mathematical problem solving unsing comparative judgement. International Journal of Science and Mathematics Education, 13(1), 151–177. http://doi.org/10.1007/s10763-013-9497-6
- Petticrew, M., & Roberts, H. (2006). Systematic reviews in the social sciences: A practical guide. Oxford, U.K.: Blackwell Publishing.
- Pollitt, A. (2012a). Comparative judgement for assessment. International Journal of Technology and Design Education, 22(2), 157–170. http://doi.org/10.1007/s10798-011-9189-x
- Pollitt, A. (2012b). The method of Adaptive Comparative Judgement. Assessment in Education: Principles, Policy & Practice, 19(3), 281–300. http://doi.org/10.1080/0969594X.2012.665354
- Pollitt, A., & Murray, N.L. (1995). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), Studies in language testing 3: Performance testing, cognition and assessment (pp. 74–91). Cambridge, U.K.: Cambridge University Press.

F. Developing a framework for user generated assessment

Mark Frazer and Sarah Gott (CEM, University of Durham, United Kingdom)

Recent assessment policy changes in some jurisdictions have led school leaders and teachers to look for new and alternative methods of assessing and monitoring their pupils. Advances in educational research point to the need to reconsider the currently accepted assessment paradigm.

The emergence of the demand for user generated content, or 'crowd sourced' assessment material, presents a novel set of considerations. Ensuring that these new paradigms are fit for purpose is a challenge facing the assessment community. There are few accessible frameworks to support practitioners, often non-specialists in assessment, who want to create their own items or systems.

The creation of assessment materials using a crowd sourced methodology has the potential to deliver multiple benefits to developers and practitioners alike. Advantages would include the capability for individuals and organisations to draw upon the skills and opinions of experts; to compile and develop item banks in reduced timeframes; and to collaborate with colleagues, sharing resources and developing assessment knowledge and understanding.

Underlying all assessment development is the need to define the range of functions that an assessment, or programme of assessment, should perform. Key assessment principles should provide the context within which the range of functions resides. A predetermined understanding of the interpretations to be drawn ensures that the quality of the product is fit for the purpose intended. Numerous attempts at defining the purpose and principles of assessment have been made (for example Schuwirth et al., 2011; Van Der Vleuten, 1996). How these relate specifically to crowd sourced assessments presents an interesting enigma yet to be considered. Furthermore, any developments must take into account the validity of the interpretations which are to be drawn; argumentation through discourse forms an integral part of this process.

The framework which the authors hope to develop is intended to provide a structure for the process of developing, using and contributing to crowd sourced assessments. Development principles will include validity argumentation, the creation of practical guidelines for developers and teachers, implementation advice, reporting and feedback and ongoing reviews to ensure the quality of the procedures and the product. The process will also begin to establish a common vocabulary related to crowd sourced assessments. Through an iterative process the framework will ultimately incorporate benchmarking and quality improvement.

The authors propose to draw on the knowledge and expertise of the international assessment community in developing a framework for crowd sourced assessments. The proposed discussion will form the first step in the exploration of the theoretical elements of the framework. Emerging themes, concepts and vocabulary from the discussion will contribute towards the content of the framework and future development in this area. Publications based on the development of the framework are anticipated, with credit given to all contributions.

The following research questions are to be presented to the group for discussion.
1) What are the fundamental elements of a framework for crowd sourced assessments to ensure the quality of both the process and product?
2) How can we mitigate against the misuse of interpretations made from crowd sourced assessments?

References

- Schuwirth, L., Colliver, J., Gruppen, L., Kreiter, C., Mennin, S., Onishi, H., Pangaro, L., Ringsted, C., Swanson, D., Van Der Vleuten, C. (2011). Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. Medical teacher, 33(3), 224-233.
- Van Der Vleuten, C.P.M. (1996). The assessment of professional competence: developments, research and practical implications. Advances in Health Sciences Education, 1(1), 41-67.

G. The effect of teaching referencing practices on student attitudes towards cheating

Rebecca Hamer and Tamsin Burbidge (International Baccalaureate, The Netherlands)

With the growth of technology in education, opportunities increase for students to engage in behaviours associated with cheating and plagiarism, but which students themselves may perceive as normal school behaviours (Baluena & Lamela, 2015) or a legitimate way to help their friends especially in international education (e.g. Winrow, 2015). Lack of clarity on what constitutes cheating or plagiarism in both students and teachers may contribute to the prevalence of academic misconduct, possibly resulting in more unethical behaviours in the workplace and reputational damage to educational institutes (Balbuena & Lamela, 2015). Research seems to indicate that active teaching of academic integrity practices, enculturation of these practices over time and role modelling of these practices by teachers influences student perceptions and attitudes (e.g. Bretag et al., 2014), although it is not always clear what and how it should be taught (Löfström et al. 2014).

The International Baccalaureate (IB) offers the Diploma Programme (DP), a two-year pre-university level certificate, in 148 countries worldwide. It requires all schools offering the DP to have an Academic Honesty school policy and to engender a culture of academic integrity in their graduates and staff, enforcing this requirement at the exam level expecting schools to teach and uphold the spirit of this policy so that students are well prepared. Ubiquitous use of technology and the internet in the learning-teaching environment complicate and increase the effort to go into monitoring any breaches and increases the need to support schools in how they implement and uphold this policy. In 2015 IB undertook a large scale survey of IB students, teachers and school administrators (i.e. DP coordinators), triangulating attitudes, teaching practices and experiences of these three respondent groups in the DP, and creating a link to academic integrity in higher education by using adaptations of well-established questionnaire items.

Comparing the three response groups, students more often felt some behaviors were not or only minor cheating, whilst coordinators were far stricter. The majority of IBDP students, teachers and coordinators agree that submitting other people's work as your own, e.g. by copying from another student (during exam or otherwise), or indeed submitting a paper obtained from internet as your own is serious cheating. Accepting or providing help when submitting work, e.g. helping someone cheat at an exam or submitting work on which parents did most of the work, was seen as a slightly less serious form of cheating by all, with small numbers of students and teachers indicating it was not cheating. Behaviours traditionally associated with plagiarism were only fourth in the order of perceived severity of cheating. Knowledge of and access to the school´s academic honesty policy significantly reduced the number of students perceiving these behaviours as not cheating. This trend is observed for teachers but to a lesser extent. Training in reference practices such as correctly referencing sentences or texts (translated) from the internet or other sources also reduced the number of students that felt behaviours were not, or only minor cheating, whilst training in social media use and self-citation related to referencing is uncommon.

References

- Baluena, S.E. & Lamela, R.A. (2015). Prevalence, Motives, and Views of Academic Dishonesty in Higher Education. Asia Pacific Journal of Multidisciplinary Research, 3(2), 69-75.
- Bretag, T., Mahmud, S., Wallace, M., Walker, R., McGowan, U. et al. (2014). 'Teach us how to do it properly!' An Australian academic integrity student survey. Studies in Higher Education, 39(7), 1150-1169.
- Löfström, E., Trotman, T., Funari, M. & Shapard, K. (2015). Who teaches academic integrity and how to they teach it? Higher Education, 69, 435-448. DOI 10.1007/s10734-014-9784-3.
- Winrow, A.R. (2015). Academic Integrity and the Heterogeneous Student Body. Global Education Journal, 2.

H. Investigating students' Perception on Rubric-oriented Assessment in the Micro Context of Russian University

Olga Mironova (Nizhny Novgorod Linguistics University, Russia)

The paper focuses on different kinds of assessment tools in FL learning used in the context of Russian Universities. The task of implementing best practices are explored. Since the idea of student-centered education is widely spread in Russia it is crucial to compare curriculum standards and the outcomes to be measurable and applicable. The thorough analysis of Russian university standards content showed that the outcomes in FL can be reflected in critical thinking, diverse communication skills, problem solving, reasoning, and decision making. But the main question which has so far not been solved is how to assess all these skills in the dimensions of language learning. This research is an attempt to find an effective tool for evaluating educational outcomes in FL classes at tertiary level. Rubric-based assessment provides the teacher and student with a tool for a conducting meaningful, criterion-referenced assessment. Such assessment provides the learner with a clear picture of their learning and of areas for potential growth. Rubrics help to limit teacher subjectivity by encouraging thoughtful representation of the learning goals for the lesson or unit in advance of the instruction.

The authors deal with criterion-referenced assessment. The basic idea is that student performance is compared to a standard of achievement rather than to other students. A rubric is a tool to help a teacher compare the progress and achievement of their students to the desired outcomes for the instructional unit. As one of the most popular assessment tools in educational programs is the rubric the given study is based on the exploring students` perception using rubrics for assessing the outcomes in FL classes.

The context of the study was Nizhny Novgorod Linguistics University and Samara University, Russia, and the participants in this study were 16 postgraduate students taking EFL course. The classes meet two times per week for two hours each session during 15 weeks. Postgraduate students who have taken part in the study are from different academic areas (History, Sociology, Psychology, and Pedagogy). The learners are studying English for passing their postgraduate (EAP) exam and improving their English skills for sharing their research through papers, presentations and communication. There are four female and six male students whose age ranges from approximately 23-39. The proficiency level is B1 – 14 students and B2 – 2 students

(according to their placement test in the beginning of the course). The learners are from Russia, they speak Russian as their first language. Students are all literate in their first language.
The main target of the course is to teach students to be a part of the global scientific community through academic communication by analyzing and discussing scientific papers in their area. At the end of the course the learners are expected to to write a literature review that is a text written to consider the critical points of current knowledge including substantive findings, theoretical and methodological contributions to a particular topic in the field of postgraduate students' research. The activities of the EAP classes were selected to address the course objectives (from the course syllabus). In addition, a variety of original instructional activities are devised, in which postgraduate students are focused on improving their academic English skills many of which involve authentic reading, interpreting materials, such as articles from journals, monographs. Most student output in the classroom consists of presentations to display their work with authentic literature and their research. Student performance is assessed both formatively through homework assignments and performance during classroom tasks, and summatively through unit exams, a midterm, and a final exam. To pass this 3 credit course students must achieve an overall grade of 70%.

I. How is ICAEW, a global chartered accountancy body, replicating the realities of the workplace in professional exams without compromising quality and rigour?

Mike Green (RM Results, United Kingdom)

Recently, global chartered accountancy body ICAEW surveyed over 100,000 members and key stakeholders to help determine the future shape of its ACA qualification. A key finding was that students and employers expected not only alignment between syllabus, tuition and assessment – but between these and the day-to-day realities of their working lives. Employers highlighted a limited use of pen and paper in the office and called for professional examinations to reflect this.

Technology plays a large part in the workplace: students are familiar with using technology to write reports, create spreadsheets, tables and undertake computations. ICAEW's exams need to reflect the activities students are already undertaking as part of their ACA training.

ICAEW began their move to computer-based exams in 2007 with the Certificate Level assessments and were keen to follow this success for the ACA Qualification.

This paper will set out the steps taken by ICAEW, supported by technology partner RM Results, from recognising the requirements of its membership and stakeholders, to fully implementing end-to-end computer-based examinations (CBE).

The two organisations will explain the phased stages of the procurement and implementation processes, designed to limit overall risk and exposure.

The paper will delineate other factors supporting a move to CBE, including enhanced security, increased flexibility and a drive for competitive advantage. Above all it will illustrate how a move to CBE has been achieved for ICAEW without any reduction in the current quality and rigour of the ACA qualification. Rather, it will set out how the introduction of computer technology will support the evolution of ICAEW's assessment methodology.

RM Results will explain how the 'mark anything, anywhere' capabilities of its RM Assessor e-marking platform have supported the transition from paper to electronic exams for ICAEW without interrupting the human marking process which ensures high-quality assessment.
In conclusion, ICAEW and RM Results will explain how the new integrated CBE and e-marking platform will allow future developments in technology to be incorporated as part of the continuing evolution of the ACA qualification.

J. Initiative of criteria-based assessment system implementation in Kazakhstan

Olga Mozhayeva, Aidana Shilibekova and Aliya Mustafina (AEO 'Nazarbayev Intellectual Schools', Kazakhstan)

Today assessment is no longer considered to be the element of control. It becomes one of the essential directions of strategic development of countries' educational policy.

Kazakhstan is not an exception to the rule. In 2014 OECD presented the 'Review of National Politics for Education' report, where it strongly criticized the current assessment system of Kazakhstan by indicating the uncertainty in the learning objectives, inappropriate use of different types of grades, awarding of over-estimated grades by teachers, etc.

The government understands the importance of rethinking of the current assessment system, which is based on the comparison of student's results with other students' achievements (so-called 'norm'). One of the main elements of this assessment system is a 5-point grading scale, where there are no clearly defined rules of grading and the arbitrariness in establishing the norms and criteria of assessment. Student's final grade is calculated as an arithmetic mean of all received grades (grades have equal weight).

As a result, in 2015-2016 academic year the government initiated the implementation of the Updated Content of Secondary Education Project in the Republic of Kazakhstan. The program acquisition is supposed to be assessed through concrete results, which are reflected in student's knowledge, skills and experience. Therefore, one of the prioritized directions of this project is considered to be the introduction of criteria-based assessment system of students' academic achievements. The introduced criteria-based assessment system is oriented on the development of student, the increase of student's motivation toward education, and the insurance of assessment objectivity, continuity and reliability. In order to satisfy these requirements, clear and measurable assessment criteria, which can be understood by students and their parents, have been set up, and the appropriate mechanisms of assessment in a teacher's practice have been defined. The continuity of assessment will be defined by summative accumulation of points. The determination of student's opportunities, the identification of problem areas and the consummation of the highest results will be provided by formative assessment.

New initiative raised many questions among the Kazakhstani population. Therefore, the authors have decided to undertake an additional research on the following topic. The article presents empirical results of two monitorings (at entrance and output), which took place among the 1st grade students (more than 5 000 students) of 46 public secondary schools in Kazakhstan. Criteria-based assessment system proved its effectiveness during the approbation among public secondary schools' students. In addition, the comparative analysis of the current and new assessment systems will be presented. The article will include the review of: the current assessment practice, policy regarding the assessment of students' academic achievements, taken actions in terms of initiative implementation (criteria-based assessment system) and future perspective directions.

K. You read on screen, I read on paper – Are we reading the same texts?

Ragnhild Engdal Jensen (University of Oslo, Norway)

Traditionally the distinction between paper-based and digital texts has been used in the PISA framework, showing that the medium, in which texts are presented, is crucial for the categorization. In the PISA 2015 framework this distinction is no longer found, due to the change in delivery mode. Although it is emphasized that this change implies a breach with previous assessments, it is argued that the presentation of print texts on screen in an assessment situation is no longer a violation of authenticity, as electronic devices to an increasing extent are being used for different reading purposes (OECD 2013). Simultaneously the change provoked a new distinction between fixed and dynamic texts, reflecting whether or not the reader has the opportunity to interact with the text, and the extent to which the texts requires the reader to use navigation tools. Some argue that texts consumed on screen are in essence moveable, dynamic and changeable, and scarce existing research do show that students who read texts on paper perform significantly better than those who read texts on screen (Mangen et.al., 2013), perceiving the latter more challenging even if the texts are the same (Murphy & Imrie, 2003).

In Norway, the national reading test is also transferred to screen from 2016. This shift from paper to screen has both practical and logistical reasons, but there is a concern whether the test continues to measure the same underlying concept of reading as before. Since both PISA and the national tests have undergone an extensive change, which might affect the way students read, solve tasks, and gain results, we explore (1) to what extent the paper-based and on-screen reading assessments measure the same underlying construct, and (2) if delivery mode influences students' reading comprehension.

The study utilizes data collected in PISA 2015 and two pilot studies from the Norwegian national reading test. The two assessments share many characteristics, due to which the PISA framework is used as the starting point for categorizing texts and items. Item response theory will be used to compare the paper-version against the screen-version of the assessments. Furthermore, these findings will also be used to identify characteristics of students who do not perform as expected when changing the test delivery mode and as a basis for development of a parallel test by utilizing items for which a difference between the two media has been detected. In this way and by means of eye tracking procedures we will be able to follow and compare the same students as they read on paper and on screen.

The results will provide a deeper understanding of whether or not paper-based and on-screen reading assessments measure the same construct. As the transition of assessments is driven by technology rather than knowledge or pedagogical considerations, results from this study will provide important insight to policy makers of what the assessment entails and what kind of additional preparations that need to be made to ensure quality assurance. For teachers and school leaders the results could lead to increased awareness in terms of challenges related to the reading of different text modes, and provide a good basis for improvement of reading instruction across subjects.

References

- Mangen, A.,Walgermo, B.R. and Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. International Journal of Educational Research, 58 (2013), 61-68.
- Murphy, P.M & Imrie, A. (2003). Implementing computers in a reading classroom. Studies in Linguistics and Language Education of the Research Institute of Language Studies and Language Education, Kanda University of Interantional Studies, 14, 123-166.
- OECD (2013), PISA 2015 Draft Reading Literacy Framework. Paris: OECD Publishing.

L. A site of tension: the complex case of GCSE English speaking and listening

Ruth Johnson (AQA, United Kingdom)

Existing research suggests that high-stakes testing leads to a narrowing of the curriculum as teachers focus on teaching to the test (Berliner, 2011). In the context of England the GCSE in English is extremely high stakes for student, teacher and school. It is a crucial component of governmental accountability measures (DfE, 2014), creating enormous pressure for teachers (Perryman, Ball, Maguire and Braun, 2011). In focusing on this high-stakes test, this poster sheds further light on the tensions between teaching and learning and the priorities occasioned by accountability measures.

This poster reports the findings of a qualitative study which explores students', teachers' and examiners' conceptualisations of and responses to the English GCSE. While the research project as a whole addressed the different components of the assessment, this paper focuses on the speaking and listening assessment, which constituted 20% of the overall assessment when the research was conducted. A total of 24 semi-structured interviews were conducted with students and teachers across three schools and with senior examiners. The samples of schools and the students within those schools were purposive and were selected to ensure a diverse range of voices was represented.

The study reported on here found that the senior examiners interviewed conceptualised the speaking and listening assessment idealistically, as a site of empowerment for students from all social and cultural contexts. Students and teachers however, in most cases, conceptualised speaking and listening as discrete assessment tasks. In very few cases was speaking and listening understood as a site of teaching or learning. Given the wealth of research that emphasises a causal relationship between high-stakes tests and teaching to the test, this is initially surprising: there are several possible explanations. It is possible, given the ephemeral nature of the assessment evidence, that teachers haven't focused on speaking and listening because they thought they could 'get away with' not doing so. An alternative explanation, which is supported by analysis of the qualitative data, is that teachers don't focus on teaching speaking and listening skills because they have an implicit understanding that those skills are inherently connected with the social and cultural backgrounds of their students. The teachers interviewed expressed a belief that whatever they did, they couldn't equip their students with the skills necessary to achieve the highest marks on the assessment because of the focus of the assessment on Standard English and sophisticated vocabulary. Responding to this understanding, the teachers focussed instead their attention on the parts of the test where they felt they could make a tangible difference.

Many of the students also saw speaking and listening as a site of discomfort and anxiety. They felt threatened by it in a way which wasn't seen at all in relation to the assessment of reading and writing. This poster concludes by suggesting that limiting speaking and listening to an assessment task in the margins of the curriculum does students from less entitled backgrounds no favours. Explicit teaching of speaking and listening would enable students to access not just the assessment but an important set of skills for life.

References

- Berliner, D. (2011) Rational responses to high stakes testing: the case of curriculum narrowing and the harm that follows, Cambridge Journal of Education, 41:3, 287 – 302.
- DfE (2014) Statement of Intent [online] (London, DfE), available at http://www.education.gov.uk/schools/performance/download/Statement_of_Intent_2014.pdf.
- Perryman, J., Ball, S., Maguire, M. and Braun, A. (2011) Life in the Pressure Cooker – School league tables and the English and mathematics teachers' responses to accountability in a results-driven era, British Journal of Educational Studies, 59:2, 179 – 195.

M. The Internationalization of Higher Education: assessing university staff

Victoria Levchenko (Samara University, Russia)

Globalization is the context of economic and academic trends that are a part of the reality of the 21st century. The motivations for internationalization include knowledge and language acquisition, enhancing the curriculum with international content. Specific initiatives such as cross-border collaborative arrangements, programs for international students, establishing English-medium programs and degrees, etc. have been put into place as part of internationalization. Many higher education institutions around the world have internationalized their degrees and programs, and they have established foreign branch campuses to provide their intellectual resources in other countries. However there is a lack of empirical evidence and conceptual understanding of how universities should adopt the international initiatives. Accumulation a lot of experience, including international, in developing courses, syllabuses and assessment criteria helps to develop a conceptual framework of the university English language leaning and teaching policy with definite outcomes and assessment criteria that will effectively prepare faculty to teach in the context of higher education globalization.

The major goal of this research is to develop a sound methodology for creating English language learning and teaching environment as a prerequisite for successful internationalization. The first stage of the project focuses on creating standards that will explicitly outline the abilities and levels of performance expected from a teacher. The statements should be written in a clear language, it should not be too broad or too narrow, and should not mention the specific task a respondent will be required to perform. The standards must be observable and measurable.
The next stage is codification of standards which means that all the standards are written in a consistent style and all definitions should be validated.

The most crucial part of the research is the development of assessment criteria for the university staff in order to help them become more efficient in their teaching and innovative in their work. The methodology of the English language policy is going to provide a framework for internationalization because it will target the competencies university teachers require in an internationalized environment.

The outcomes of the research are:
1. Valid and reliable methodology of the English language Policy
2. A framework for assessing competencies of the teaching staff
3. A list of strategies that would prepare faculty to teach cross-culturally.

The introduction of the methodology, new assessment bodies, recruitment criteria and institutional strategies in the field of training, monitoring and motivation of the faculty will contribute to a successful process of internationalization and integration into higher education global environment.

N. Provision of feedback in L2 exam classes in Cyprus

Dina Tsagari and George Michaeloudes (University of Cyprus, Cyprus)

The language testing field has placed a great importance on the impact of examinations, in particular on the ‘washback effect’ of high-stakes language tests on teaching and learning (Alderson & Wall, 1993). However, the scope of research of test washback on teaching and learning accumulated over the past two decades has been somewhat limited (Tsagari & Cheng, forthcoming). Researchers have considered the influence of tests on teacher perceptions and attitudes or instructional aspects such as ‘methods’ or ‘tasks’ and ‘activities’ but very little research has been undertaken in teacher-student interaction and feedback provision in exam preparation, despite the importance placed on it in second language (L2) acquisition (Fujii & Mackey, 2009; Mackey, 2012). Such issues have recently problematized language assessment (Harding, 2014; Harris, Smith, & Harris, 2011; Wall & Taylor, 2014) while feedback has become the focus of emerging theories and discussions in the field (Hasselgreen, 2012; Tsagari, 2014b; Turner & Purpura, 2016).

The purpose of the present study was to investigate instances of classroom interaction where teachers provide students with feedback to support and scaffold the learning of English in exam preparation contexts. The study focused on Cambridge English: First (Cambridge English Language Assessment), to gain a deeper insight into the classroom realities of exam preparation.
The analysis of the results showed that teachers introduced a number of implicit and explicit strategies to scaffold learners towards achieving the main goal of their exam classes, e.g. successful preparation for the exam. Nevertheless, feedback was restrictive, usually provided through short-term rewards rather than detailed analysis of problems, with little explanation of

what students' strengths and mistakes and how improvement could be made or maintained. Also, the analysis showed an overemphasis on grammar, frequent use of L1 to provide explanations and advice, and rapid rhythm of checking answers and very little space for authentic use of the language. Nevertheless, the findings of the study have pedagogical and research implications. For instance, teachers in the present context could benefit from in-service training that can inform them about effective ways to prepare learners for high-stake examinations (e.g. formative feedback) without losing sight of their students' linguistic and pedagogic needs. Also future research should explore teachers' (and pupils') perceptions, understanding and practices of particular feedback provision e.g. through stimulated recall interviews and larger number of participants. This will reveal the extent to which teachers and pupils share a common understanding of the nature and purpose of feedback input as part of their larger formative assessment processes and the way in which the interaction outcomes of feedback variations can improve learning.

O. Computerized standard setting using the Data-Driven Direct Consensus (3DC)

Jesse Koops, Remco Feskens and Frans Kleintjes (Cito, The Netherlands)

Among the best-known examples are the Angoff procedure, the Bookmark procedure and the Direct Consensus procedure. These procedures have their strengths and weaknesses. In the present study, the strengths of the aforementioned standard setting procedures were brought together in a new one: the Data-Driven Direct Consensus (3DC) procedure. The 3DC procedure divides the full test into a number of clusters and uses empirical data to relate the scores of the clusters to the scores of the full test. The relationships between the clusters and the full test are presented to the panellists on a specially designed assessment form. Panellists are asked to use the assessment form to indicate the score that students would be expected to achieve in each cluster if they were exactly on the borderline of the selected mastery level. Because of the design of the assessment form, the assessment is easily allowed to be based on both content information and empirical data.

The 3DC procedure has several advantages. First, the cognitive burden to complete the task can be reduced, especially if the clusters are relatively small. The subject-area experts are then still able to form an opinion about the minimally expected performance on the entire cluster and the focus on clusters, moreover, appears to be a cognitively easier task than the focus on single test items as is common for the Angoff and the Bookmark procedures. Contrary to the Direct Consensus method, which also divides a test into clusters, the 3DC procedure presents empirical information about the relative difficulty of clusters to the panellists. This can be considered a second advantage of the methodology: the empirical information can assist the panellists in the evaluation of the test materials. Third, the 3DC procedure provides flexibility in the use of item types and statistical models. A variety of item types from, for example, multiple-choice to constructed-response questions can be included in the standard setting and basically any predication model can be used to construct the assessment form. Alternative procedures usually do not offer this kind of flexibility. Finally 3DC offers many opportunities for evaluating the correspondence within and between subject-area experts. As subject-area experts are asked to set a passing score on several clusters, it is possible, for instance, to evaluate rating consistency across the clusters.

An important development in the use of the 3DC procedure is the construction of a digital platform which can be used to conduct the standard setting. To date, most 3DC standard setting procedures were conducted using paper prints of the assessment form. Panellists used a ballpoint pen to indicate their judgement on the assessment form and after each assessment round the ratings were collected and filled in by the moderator into an Excel summary file. These steps can now also be performed using the digital platform. Each panellist uses his or her own laptop and the assessment form appears on the laptop of each panellist. The panellists then indicate their judgements on the digital assessment form and by one simple action the moderator can centrally collect the judgements of all the panellists. The summary file will also be filled in immediately. The digital platform has already been tested and used in several standard setting conferences. Evaluations from both the panellist and the moderators were positive; the platform makes the procedure much more user-friendly. A free copy of the application can be downloaded from www.cito.com/3DC

Reference

- The Data-Driven Direct Consensus (3DC) Procedure: A New Approach to Standard Setting
  Jos Keuning, J. Hendrik Straat, Remco C.W. Feskens and Karen Keune
  Cito, Institute for Educational Measurement, The Netherlands

P. What make PISA items more difficult for students with minority background? Analysing the effects of item interactivity and response format in a computer-based assessment of scientific literacy

Nani Teig (University of Oslo, Norway)

A growing number of assessments, including The Programme for International Student Assessment (PISA), have taken advantage of technological innovations by adopting new assessment formats which include interactive and dynamic materials (e.g. simulations). Interactive items using simulated real-world problems are considered beneficial in assessing complex inquiry skills. However, the results of previous research regarding the effect of item interactivity on students' performance have been mixed. Students who used an interactive system outperformed those using static systems (Quellmalz, Timms, Silberglitt, & Buckley, 2012). Second-language learners also performed significantly higher on interactive, simulation-based science assessments. Conversely, DeBoer et al. (2014) found that students' performance decreased when assessment items changed from static to interactive.

Instead of assuming that interactive items are better than static items, or vice versa, we should take into account not only how the items are presented but also what type of responses are required to answer them. Students might perform better on interactive than static tasks when the response formats are simple (i.e., multiple-choice). When the response format is more complicated, such as a combination of multiple-choice and written response, their performance on interactive tasks might decrease. Therefore, this project focuses on understanding the effect of item interactivity (static vs. interactive) and the complexity of response format (i. multiple-choice, ii. multiple-choice and simulated response, iii. multiple-choice and written response, iv. multiple-choice, simulated response, and written response) with the following research question (RQ):

1. To what extent do item interactivity and response format affect item difficulty?

   Substantial gaps in performance between majority and minority students are often found in many countries. Students whose first language differs from the language of assessment face severe challenges related to reading and writing load. Little research has been done to understand the factors that influence performance differences at the item level, especially on the items that employ simulation. Interactive items that maximize the use of simulated phenomenon for scientific investigation might reduce the reading load, making the items easier for minority-language students. However, item difficulty might increase when the response format requires a high writing load (i.e., written response). To understand the performance gap between different student groups at the item level, the following question will be explored:

2. Can item interactivity and response format account for differences in item difficulty between students with majority and minority backgrounds?

Method

This study uses data from the Norwegian PISA 2015 sample. Item response theory will be applied to analyse the data from a multiple matrix-sampled assessment. An explanatory item response theory approach is adopted to disentangle the effects of item and person characteristics simultaneously. This approach allows for studying the effects on item difficulty directly.

Expected Outcomes

For RQ-1, PISA items become easier when presented in an interactive and simulated system. After controlling for cognitive demand, competency, and type of knowledge required to solve the items, item difficulty increases when the response format is more complex. Interactive items and simple response formats reduce the performance gap and provide equal opportunities for different groups of students (RQ-2).

References

- DeBoer, G. E., Quellmalz, E. S., Davenport, J. L., Timms, M. J., HerrmannğAbell, C. F., Buckley, B. C., Flanagan, J. C. (2014). Comparing three online testing modalities: Using static, active, and interactive online testing modalities to assess middle school students' understanding of fundamental ideas and use of inquiry skills related to ecosystems. Journal of Research in Science Teaching, 51(4), 523-554.
- Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. Journal of Research in Science Teaching, 49(3), 363-393.

AEA-Europe | Association for Educational Assessment - Europe

AEA-Europe | Association for Educational Assessment - Europe

AEA Europe | About AEA-Europe

AEA-Europe is a membership organisation set up in 2000 to support and develop the assessment community throughout the whole of Europe.

AEA-Europe offers its members a range of opportunities to network with each other, sharing news, debate and research. At institution level, the Association provides a forum for international liaison and co-operation.

AEA-Europe members have access to:

1. Professional development opportunities
   - Accreditation scheme- recognition of experience, knowledge and expertise in asessment at Practitioner and Fellow levels
2. Discussion and debate opportunities via our regular online newsletter
3. Our annual autumn conference
   - Pre-conference workshops
   - Keynote presentations on topical issues in assessment
   - Discussions and debates
   - Social programme

And each year a new European city to get to knowl

For more about AEA-Europe and how to join, visit http:/lwww.aea-europe.net/

AEA-Europe | The Council

President | Guri Nortvedt
Department of Teacher Education and School Research, University of Oslo, Norway
g.a.nortvedt@ils.uio.no

Vice President | Thierry Rocher
Directorate for Assessment, Forecasting and Performance (DEPP), France
thierry.rocher@education.gouv.fr

Executive Secretary | Alex Scharaschkin
AQA, United Kingdom
AScharaschkin@aqa.org.uk

Treasurer | Cor Sluijter
Cito, Institute for Educational Measurement, The Netherlands
Cor.sluijter@cito.nl

Council member | Iasonas Lamprianou
Department of Social and Political Sciences, University of Cyprus, Cyprus
iasonas@ucy.ac.cy

Council member | Gill Stewart
SQA, United Kingdom
Gill.Stewart@sqa.org.uk

Council member | Antonella Poce*
Roma Tre University, Italy
apoce@uniroma3.it

*Antonella Poce's term ended September 2016.

AEA-Europe | Publications Committee

The AEA-Europe Publications Committee aims to share the work of the Association more widely, involving more of the membership in the Association's activities, facilitating contacts between members, and initiating publications of relevance to members. From 2016 committee members are:

- Gill Stewart, SQA (United Kingdom)
- Jeanne-Marie Ryan, Oxford University (United Kingdom)
- Anastassia Voronina, INNOVE (Estonia)
- Lesley Wiseman, Independent educational consultant (United Kingdom)
- Daniel Xerri, University of Malta (Malta)
- Amina Afif, Luxembourg Government (Luxembourg) (newsletter editor)

AEA-Europe | Professional Development Committee

The broad objective of the AEA-E Professional Development Committee is to develop initiatives that support the professional development of the members of the Association, and to organise the professional accreditation programme. The Professional Development Committee (PDC) was established during the annual AEA-Europe Conference in Paris in 2013.

Members of the PDC are:

- Antonella Poce (University Roma 3, Italy) – Council member and Committee Chair until June 2016;
- Stuart Shaw (Cambridge International Examinations-CIE, Cambridge Assessment, UK) until June 2016
- Bas Hemker (Cito, The Netherlands)
- Andrew Boyle (Alphaplus Consultancy, UK)
- Stéphanie Berger (University of Zurich, Switzerland)

AEA-Europe | Limassol Conference Organising Committee

- Iasonas Lamprianou (University of Cyprus)
- Thierry Rocher (DEPP)
- George MacBride (University of Glasgow)
- Michalis Michaelides (University of Cyprus)
- Elena Papanastasiou (University of Nicosia)
- Cor Sluijter (Cito)
- Dina Tsagari (University of Cyprus)

AEA Europe | Limassol Conference Scientific Programme Committee

- Sarah Maughan (AlphaPlus)
- Guri A. Nortvedt (University of Oslo)
- Michalis Michaelides (University of Cyprus)
- Andrej Novik (SCIO)
- Elena Papanastasiou (University of Nicosia)
- Gill Stewart (SQA)

AEA Europe | Review Panel

The Council is very grateful for the contribution of all members of the review panel:

- Alex Scharaschkin, AQA
- Andrej Novik, SCIO
- Angela Verschoor, Cito
- Anton Béguin, Cito
- Antonella Poce, Universita Roma TRE
- Ayesha Ahmed, University of Cambridge
- Bernard Veldkamp, University of Twente
- Carolyn Hutchinson, University of Glasgow
- Christina Wikström, Umeå University
- Cor Sluijters, Cito
- Daryl Stevens, AQA
- Dina Tsagari, Department of English Studies, University of Cyprus
- Elena Papanastasiou, University of Nicosia
- Ernie Spencer, University of Glasgow
- Fabienne van der Kleij, Australian Catholic University
- Filio Constantinou, Cambridge Assessment, University of Cambridge
- Frans Kleintjes, Cito
- George MacBride, University of Glasgow
- Gill Stewart, SQA
- Gordon Stobart, IoE
- Grace Grima, Pearson UK
- Guri A. Nortvedt, University of Oslo
- Iasonas Lamprianou, University of Cyprus
- Jana Straková, Institute for Research and Development of Education
- Jannette Elwood, Queen's University Belfast
- Jaroslava Simonova, Charles University in Prague
- Louise Hayward, University of Glasgow
- Maria Teresa Florez Petour, Pedagogical Studies Department, University of Chile
- Michalis Michaelides, University of Cyprus
- Natalie Usher, University of Oxford
- Paul Newton, Ofqual
- Roger Murphy, Consultant
- Rose Clesham, Pearson
- Sandra Johnson, Assessment Europe
- Sarah Maughan, AlphaPlus
- Steven Bakker, dutchTest
- Stuart Shaw, Cambridge Assessment
- Tandi 'clausen-May', Independent Consultant
- Theo Eggen, Cito/University of Twente
- Thierry Rocher, DEPP
- Yasmine El Masri, Oxford University Centre for Educational Assessment

AEA-Europe | The Kathleen Tattersall New Assessment Researcher Award review panel

Each year the PDC appoints a panel to review the applications that have met the Criteria for Eligibility. The 2016 panel consisted of three senior assessment researchers drawn from the Fellows of the Association. To avoid conflict of interest, no member of the review panel worked at the same institution of, supervised any of the applicants being judged or has provided them with a letter of recommendation for the award panel.

In 2016, the review panel were Paul Newton, Julie Sewell and Simon Wolming.

The 2016 Kathleen Tattersall New Researcher Award Winner is Sebastiaan de Klerk.

Notes

Notes

Notes

Notes

Sponsors

CAMBRIDGE ASSESSMENT

Springer

CYPRUS TOURISM ORGANISATION

Passion FOR SUCCESS

Austrian

CiTO

Partners

Vretta™

easy CONFERENCES

AEA-Europe | Association for Educational Assessment - Europe

Social and Political underpinnings of educational assessment: Past, present and future

The 17th Annual AEA-Europe Conference