AEA-Europe | Association for Educational Assessment - Europe

Assessment for transformation

Teaching, learning and improving educational outcomes

The 20th Annual AEA-Europe Conference

Programme | 13-16 November 2019

Lisbon, Portugal

aea EUROPE

IAVE INSTITUTO DE AVALIAÇÃO EDUCATIVA, I.P.

Contents

Introduction	**2**
Programme	**6**
Poster Presentations	**30**
Open Paper Sessions	**40**
About AEA-Europe	**98**
The Council	**98**
Publications Committee	**99**
Professional Development Committee	**99**
Audit Committee	**99**
Conference Local Organising Committee	**99**
Conference Scientific Programme Committee	**99**
Review Panel	**100**
The Kathleen Tattersall New Assessment Researcher Award review panel	**101**
The Accreditation review panel	**101**

Introduction

Assessment for transformation: teaching, learning and improving educational outcomes

As we gather again for our annual conference, I have been reflecting on the theme chosen by our Portuguese colleagues that focuses on the transformative power of assessment.
As we well know, too often assessment is used as a political tool in educational reform, with specific forms of assessment being chosen and aligned with particular views of knowledge and curriculum to influence what is taught and assessed in schools. If we stop and reflect on the many changes that our own governments have made over the last few years, I am sure we can recognise assessment being used in this way in our own contexts. However, as the vast range of research that will be highlighted, debated and discussed at this conference will show, assessment can be used as a force for good; rather than it distort teaching and learning to the detriment of young people's educational experiences, it can enhance engagement as well as improve understanding and provide opportunities for all students to evidence their learning in significant and meaningful ways. Assessment as a force for good will prevail if we use it wisely in our practices and commendably in our systems and policies. To practice wisely and strategise commendably we no longer need to see assessment as separate from the learner/teacher. We no longer need to treat learners (and their teachers) as autonomous beings but work more tirelessly to understand the social and cultural mediation of learning and assessing as it is carried out in the complex contexts of our policy systems, schools and classrooms.

For me, what makes our AEA-Europe conferences so intriguing, is the diversity of context and culture from across Europe and beyond that is brought to bear on the study of assessment. This rich tapestry of culture and context must (and does) inform our research; indeed the tapestry is fundamental to how we conduct our research and as such should be embraced as part of the enterprise. Research into assessment cannot all be from positions of neutrality and scientific reification; assessment is a human endeavour, designed by humans to be conducted on and with humans and as such our research should always reflect this. Educational assessment demands that we investigate it from all aspects of our cultures; that the humanities and arts are as pertinent to understanding the transformative aspects of assessment as much as the social sciences. By going about our business of research, evaluation and assessment development with changed outlooks influenced by evidence and argument, we can truly enhance the transformative power of assessment to improve educational outcomes.

As always this conference could not take place without the hard work and commitment of many members and friends of the Association. On behalf of the AEA-Europe Council, I would like to take this opportunity to especially thank our colleagues at IAVE who are hosting us this year in their beautiful capital, Lisbon. We express our thanks to the President of IAVE, Luís Miguel Pereira dos Santos, for all his support and to the IAVE colleagues we have worked with to this year's conference off the ground, especially – Natália Nunes, Maria Borges, Amália Costa, Rui Pires, Manuel Gomes – thank you for all your endeavours; they have really gone above and beyond to make the conference a success!

I would also like to thank the members of the Conference Organising Committee, the Scientific Programme Committee, Easy Conferences (especially Kyriakos) and all our sponsors. Without your efforts and support we would not have a conference for delegates to attend – so thank you all.

Finally I would like to thank my colleagues on the AEA-Europe Executive Council for all their hard work and commitment to the cause, and their unflinching support in this, my first year as President! My deepest thanks also go to all those members who give their time during the

year on the Professional Development Committee, the Publications Committee, the e-assessment SIG and all who reviewed for us – submissions for the conference, applications for accreditation and applications for the Kathleen Tattersall New Researcher Award – and to those who have agreed to chair sessions at the conference. It is easy to say, but I really want you to know how much your efforts are appreciated. The Association and the Annual Conference would be nothing without the contributions from you its members, and for that we are always grateful.
I have no doubt it will be a great conference – transforming assessment as a force for good! Enjoy!

Jannette Elwood
President AEA-Europe

Conference Theme:
Assessment for transformation: teaching, learning and improving educational outcomes

The 2019 conference theme might be seen as a continuity of the theme for 2018 and embraces not necessarily new perspectives on approaches to student assessment but aims to emphasise assessment as a tool that can play a relevant role in transforming the ways students are taught, the ways they develop as leaners and, furthermore, that can contribute to high quality educational outcomes around the world.

We as an association would argue that assessment, either within the classroom or applied in an external context, is absolutely necessary to regulate and to inform teachers, families, students and politicians, as well as educational researchers, about the ways teaching and learning processes can improve. Assessment can be used mostly for formative purposes but it can also have summative aims, or a combination of both; it can assume the form of tests or exams, but can also be based on more creative and rich formats, e.g., portfolios, debates, reports, essays, collaborative tasks, among others. Moreover, for the past two decades, within many educational systems, we have witnessed the reshaping of assessment, through the emergence of more digital assessment contexts and platforms.

Despite all these opportunities and options, assessment remains a powerful tool to enhance quality education and improve student outcomes. While considering that the purpose and validity of assessment should be guiding influences to be followed at all times, in most educational systems we still face the effect of the dominance of testing. Many of the well-known, negative impacts of exams and summative assessment in classroom environments still remain, affecting the way teachers teach and students learn; valuing a mostly 'teaching to the test/learning to the test' approach that can lead to shallow learning outcomes, which in part might explain the results that some countries show in international students' assessment programmes, like PISA, PIRLS or TIMSS. This widespread approach is preventing students, especially those who are socially most vulnerable, from experiencing opportunities for optimal learning and, therefore, compromising the quality of education and the role that assessment can play, in the broadest sense, to achieving success for future generations, in a sustainable way.

Transformation must emphasise the relevance of quality and sustainable feedback. Although feedback is traditionally linked to the formative dimension of classroom assessment, it is fundamentally important that we start associating feedback with other assessment contexts. Low-stakes external assessment is one, but reporting results of summative external assessment should also be regarded as an opportunity for a positive "washback effect" insofar as the results should allow for the improvement of teaching and learning for future students cohorts. Last but not least, digital assessment should be seen, not as an end in itself, but as a way to facilitate this new approach to our schools and educational systems as a whole.

The centrality of assessment in education is something that we as educational assessment researchers recognise as a value that we must preserve. Therefore it is our common responsibility to be able to transform assessment for the better, whether we talk about its purposes, about the instruments we use or the way outcomes are reported, analysed and eventually used to (re)shape teaching and learning. We need to do so in ways that will create sustainable learning cultures that will lead to educational success. We must remember that education is an open system, in which other stakeholders, like parents and opinion makers, journalists and politicians, all play a relevant role, despite not being, for the most part, technically aware of the best options to promote educational success; very often thinking that replicating the «school» they attended some decades ago is the best solution for their children. Bearing this in mind, changes in approaches to teaching, learning and assessment must prove to be fit for purpose, show effective positive results and clearly be seen, by those key stakeholders immediately outside the educational assessment environment, as the route to be followed.

Programme

Wednesday, 13th November

8.00 -10.30 Registration

9.00 - 16.30 Pre-conference workshops, Sana Lisboa Hotel

13.00 - 14.00 Lunch

Workshop 1, Room: Castelo 9

9:00 - 16:30 **Is Assessment Fair?**

Presenters: *Isabel Nisbet[1], Stuart Shaw[2]*
[1]University of Cambridge, United Kingdom
[2]Cambridge Assessment, United Kingdom

Fairness in assessment is both complex and contentious. Assessment experts may disagree on whether scores from a particular testing program are fair. However, most assessment experts agree that fairness is a fundamental aspect of validity. As a consequence, fairness has been elevated to a greater position of prominence in the assessment literature, so much so, that it is now considered one of the three primary measurement standards that must be met to legitimise a proposed test (the other two being validity and reliability). In this workshop, we will distinguish between different uses of "fair" which have relevance to assessment. Then, we will identify some of the "lenses" used to examine fairness in assessment and suggesting a framework of questions which can be applied to lenses (measurement, legal, social justice and philosophy). Each approach will be subject to a common set of questions which will investigate whether there is an established consensus on fairness, whether that consensus should be questioned, what comprises an areas of dispute, and what the implications are for other lenses. The research will culminate in a fairness agenda for the 21st Century.

Workshop 2, Room: Castelo 8

9:00 - 16:30 **IRT in R made easy**

Presenters: *Remco Feskens[1, 2], Jesse Koops[1]*
[1]Cito, Netherlands
[2]Twente University, Netherlands

Item Response Theory (IRT) is a general statistical theory about item and test performance and how performance relates to the abilities that are measured by the items in the test. IRT provides a flexible framework which can be used to obtain comparable ability estimates even when different examinees answered different questions. For among others this reason, IRT has become the method of choice for many organizations.

In recent years, R has become the standard software platform for data manipulation, analysis and visualization. Many statistical and psychometric functions are available in R and there are several packages for doing IRT analyses. Unlike e.g. SPSS, R has no standard GUI menus, instead analysis is done by typing statements. This is a hurdle for many analysts, although programming in R is not inherently more complex than clicking buttons in SPSS.
To overcome the programming hurdle we will start with a gentle introduction in R. After that, an introduction to dexter, an R package which can be used to analyze test data using IRT and Classical Test Theory (CTT) techniques, will be given. The theoretical foundations of IRT will be concisely explained and participants will perform IRT analyses on PISA and/or their own data.

Workshop 3, Room: Castelo 6-7

9:00 - 16:30 **Innovative on-screen assessment – exploring item types and paper-to-digital transition**

Presenters: *Caroline Jongkamp[1], Rebecca Hamer[2]*
[1]Cito, Netherlands
[2]International Baccalaureate, Netherlands

Since the introduction of Computer-Based Assessment (CBA), significant progress has been made in the development of constrained or closed response CBA items. However, the development and implementation of highly interactive, dynamic assessment items aimed at assessing complex thinking skills appears challenging. In the previous AEA SIG workshop, participants explored two models classifying digital items on design characteristics. These models helped participants understand existing options but lacked the link between item type and the assessment of complex thinking skills. Using working on-screen test items from a variety of sources, participants will collaboratively explore models to classify existing computer-based assessment items by type and identify links between item types and assessment objectives. The initial transition to digital assessment often involves reformatting an existing Paper-Based Assessment (PBA) item into a digital format, a transition often presenting its own significant challenges. Participants will design one or more items for a digital environment aligned to their own field of work, experiencing the variety of choices involved in designing digital items. The SIG E-Assessment pre-conference workshop will interest anyone currently involved in preparing for the digital migration of PBA and those interested in the development of CBA items for assessing complex thinking skills.

Workshop 4, Room: Castelo 10

9:00 - 16:30 **Introduction to Differential Item Functioning (DIF) analysis with R and ShinyItemAnalysis**

Presenters: *Patricia Martinková[1, 2], Adéla Drabinová[1, 3]*
[1]Institute of Computer Science, Czech Academy of Sciences, Czech Republic
[2]Faculty of Education, Charles University, Czech Republic
[3]Faculty of Mathematics and Physics, Charles University, Czech Republic

Differential Item Functioning (DIF) analysis is an analytic method useful for identifying potentially biased items in assessments. While simply comparing two groups' total scores can lead to incorrect conclusions about test fairness, many DIF detection methods have been proposed in the past, those based on total scores as well as those based on Item Response Theory (IRT) models (Martinková, Drabinová et al., 2017).

The workshop will offer an introduction into differential item functioning (DIF) detection from a practical point of view. We will introduce the mostly used DIF detection methods with their pros and cons and we will focus on their application in practice on real data examples. The free statistical software R and its packages difNLR, difR, deltaPlotR, and mirt will be used throughout the sessions. The ShinyItemAnalysis package will provide interactive user-friendly interface helpful for those who are new to R.

Workshop 5, Room: Castelo 4-5

9:00 - 16:30 **Developed selected response test items**

Presenter: *Ezekiel Sweiry[1]*
[1]AQA, United Kingdom

The purpose of this workshop is to present and discuss guidance on developing Selected Response (SR) items. The guidance is based on a synthesis of available research literature on SR item writing, relevant aspects of cognitive psychology (including models of language comprehension and working memory capacity) and the presenter's own experience of high-stakes test development across primary and secondary education in the UK.

Guidelines will focus on a range of issues including language accessibility, the central role played by distractors in affecting the difficulty and validity of SR items, unintentional cues that can betray the correct answers, and the assessment of higher-order skills. While most SR guidelines and research are based specifically on conventional multiple choice questions , this workshop will address the full range of SR item types. Structural differences between different SR item types can influence their difficulty and proneness to different validity issues.

The workshop will also consider, the use of SR e-assessment item types, the use of SR items in diagnostic assessments, and how evidence from item trialling can be used to identify problematic items, in particular through the use of distractor analysis.

18.30 -19.00 **Welcome reception for first time participants**
Location: Sana Lisboa Hotel

19.00 - 20.30 **Welcome reception for all participants**
Location: Sana Lisboa Hotel

Thursday, 14th November

8.00 - 8.45 Registration

9.00 - 9.30 **Welcome addresses**
Room: Castelo 1-2
Jannette Elwood, AEA-Europe President
Luís Santos (IAVE President)

9.30 - 10.15 **Keynote speech**
Chair: Jannette Elwood, Room: Castelo 1-2
Title: Towards a New Generation of External Assessments: Reflection on a Systematic Literature Review
Prof. Domingos Fernandes

10.15 - 10.45 Coffee break

10.45 - 11.30 **Keynote speech**
Chair: Alex Scharaschkin, Room: Castelo 1-2
Formative Assessment in Student Transition to Higher Education - A Sociocultural Perspective
Dr. Aisling Keane, Kathleen Tattersall, New Assessment Reseacher

Poster Presentations

11.30 - 12.45 **Posters**
Chair: Cor Sluijter, Room: Castelo 1-2

Poster 1 How are Abacus resources supporting transformational mathematics learning and assessment?
Ellen Barrow[1], Jennie Golding[2]
[1]Pearson Education, United Kingdom
[2]University College London Institute of Education, United Kingdom

Poster 2 System assessments for transformation: uniting forces of national assessments and the inspectorate in Flanders (Belgium)
Mieke Heyvaert[1], Ward Adelhof[2], Rianne Janssen[1]
[1]KU Leuven, Belgium
[2]Ghent University, Belgium

Poster 3 Validity considerations in digital iterations of PISA: What can process data tell us about test-taking behaviour amongst students engaging with science assessments?
Caroline McKeown[1, 2]
[1]Educational Research Centre, Ireland
[2]School of Education, Trinity College Dublin, Ireland

Poster 4	Examining the impact on student performance in Reading, Mathematics and Science in PISA, from the perspective of teachers and principals, when students are tested at different times of the year (autumn versus spring testing) *Sylvia Denner[1, 2]* *[1]Educational Research Centre, Ireland* *[2]PhD Candidate, Dublin City University School of Policy and Practice, Ireland*
Poster 5	Using a comparative judgement method to assess inter-board equivalence of reformed GCSE (9-1) Mathematics question papers *Faiza Tufail[1], David McVeigh[2]* *[1]Pearson, United Kingdom* *[2]Pearson Qualification Services, United Kingdom*
Poster 6	Developing CEFR-based Descriptors for the Assessment of Competency in Turkish as a Foreign Language *Yiğit Savuran[1]* *[1]Anadolu University, Turkey*
Poster 7	From paper-based to computer-based assessment of numeracy: some consequences for item design *Karianne Berg Bratting[1], Guri A. Nortvedt[1], Andreas Pettersen[1], Anubha Rohatgi[1]* *[1]University of Oslo, Norway*
Poster 8	The Power of Data to take smart decisions for school improvement *Senad Karavdic[1], Amina Afif[1], Graziella Losciale[1]* *[1]SCRIPT, Luxembourg*
Poster 9	Assessment for transformation at the school level? The use of parallel tests in Flanders (Belgium) *Isabel Laenen[1], Evelyn Goffin[2]* *[1]Catholic University of Leuven, Belgium* *[2]KU Leuven, Belgium*
Poster 10	In Search of Assessment that is Fit for Purpose in Character and Citizenship Education *Ng May Gay[1], Osman Abdullah[1]* *[1]Ministry of Education, Singapore*
Poster 11	Effects with learning aids on mathematical performance in vocational education in Flanders (Belgium) *Margo Vandenbroeck[1], Lien Willem[1], Rianne Janssen[1]* *[1]KU Leuven, Belgium*
Poster 12	The Effect of MathemaTIC's New Summative Assessment Format in Digital Learning Pathways for Mathematics *Arbana Miftari[1], Jacob Pucar[1]* *[1]Vretta Inc., Canada*
Poster 13	From opinion to evidence: transforming organisational culture in two Awarding Organisations *Alison Rodrigues[1], Sarah Hughes[1]* *[1]Cambridge Assessment International Education, United Kingdom*

Poster 14 From check to act: Involving stakeholders in resonating national assessment results back to educational practice
Sabine Dierick[1], Rianne Janssen[1], Koen Aesaert[1]
[1]KU Leuven, Belgium

Poster 15 The influence of differentiation on the quality of teaching gifted children
Mukhammed Mussabekov[1], Saule Vildanova[1], Nina Kashavarova[1]
[1]Center for Pedagogical Measurements under the AEO "Nazarbayev Intellectual Schools", Kazakhstan

Poster 16 Potential threats to validity through the use of extended response items
Gillian Mann[1]
[1]Scottish Qualifications Authority, United Kingdom

Poster 17 Vocational learners' perceptions on making summative assessment engaging
Vasile Rotaru[1]
[1]Qualifications Wales, United Kingdom

Poster 18 The role of motivation in performance on mathematics
Naomi Carpentier[1], Lien Willem[1]
[1]KU Leuven, Belgium

Poster 19 Higher Applications of Mathematics
Kevin Gibson[1], Martin Brown[1]
[1]Scottish Qualifications Authority, United Kingdom

Poster 20 Open questions in chemistry and physics: a creative approach to assessments of depth of knowledge and understanding of the science
Shakeh Manassian[1]
[1]Scottish Qualifications Authority, United Kingdom

Poster 21 The GCSE Mathematics Saga...
Vasu Krishnaswamy[1, 2], Jennie Golding[2]
[1]Pearson UK, United Kingdom
[2]University College London Institute of Education, United Kingdom

Poster 22 The Scandinavian legacy of resisting formal grading – paradoxes and dilemmas
Sverre Tveit[1], Lise Vikan Sandvik[2], Henning Fjørtoft[2]
[1]University of Agder, Norway
[2]NTNU Norwegian University of Science and Technology, Norway

Poster 23 Promoting job readiness for the 21st-century workplace. The empowerment of the Learning to Learn skill through the Assessment as Learning approach
Alessia Bevilacqua[1]
[1]University of Verona, Italy

Poster 24 The effect of visual-to-verbal number transcoding on mathematics achievement
Dmitrii Kholiavin[1], Diana Kaiky[1], Yulia Kuzmina[1], Galina Larina[1]
[1]National Research University Higher School of Economics, Russia

Poster 25 Building Inclusive Assessment Platforms
Luc Schomer[1], Sam Sipasseuth[1]
[1]Open Assessment Technologies, Luxembourg

12.45 - 13.45 Lunch

Open Paper Sessions

Session A: Papers 1-3 – Psychometrics I
Chair: Amina Afif, Room: Castelo 1-2

13:45 - 14:15 On IRT models for analysing high tariff items
Yaw Bimpeh[1]
[1]AQA, United Kingdom

14:15 - 14:45 Dimensionality in reading comprehension testing. An empirical validation of psychometric divisibility into reading processes
Michael Tengberg[1]
[1]Karlstad University, Sweden

14:45 - 15:15 Modeling certainty-based marking on multiple-choice items: psychometrics meets decision theory
Qian Wu[1], Rianne Janssen[1]
[1]KU Leuven, Belgium

Session B: Papers 4-6 – Educational Policy
Chair: Rebecca Hamer, Room: Castelo 9

13:45 - 14:15 Transforming assessment policy: towards a more socially just approach to policy design and enactment
María Teresa Flórez Petour[1]
[1]University of Chile, Chile

14:15 - 14:45 Making the case for assessment in educational discourses
Mary Richardson[1]
[1]UCL Institute of Education, United Kingdom

14:45 - 15:15 Critical approaches in educational assessment
Graeme Findlay[1]
[1]SQA, United Kingdom

Session C: Papers 7-9 – Test Development I
Chair: Lesley Wiseman, Room: Castelo 8

13:45 - 14:15 Spoilt for choice? Is it a good idea to let students choose which questions they answer in an exam?
Tom Bramley[1], Victoria Crisp[1]
[1]Cambridge Assessment, United Kingdom

14:15 - 14:45 White space in assessment materials – “space to think” or a “waste of space”?
Charlotte Stephenson[1], Bryan Maddox[2, 3]
[1]AQA, United Kingdom
[2]UEA, United Kingdom
[3]Assessment MicroAnalytics, United Kingdom

14:45 - 15:15 Establishing effective test length and cut-scores for formative assessment using informative Bayesian hypotheses
Hendrik Straat[1], Anton Béguin[1]
[1]Cito, Netherlands

Session D: Papers 10-11 – Higher Education
Chair: Elena Papanastasiou, Room: Castelo 6-7

13:45 - 14:15 Feedback of the external assessment of Higher Education students on the subject of Portuguese Language
Patricia Engrácia[1], João Oliveira Baptista[1]
[1]DGEEC, Portugal

14:15 - 14:45 The transformation of University admissions practices in England and its impact on A-level - are standards being maintained?
Rachel Taylor[1], Nadir Zanini[1]
[1]Ofqual, United Kingsdam

Session E: Papers 12-14 – Fairness and Social Justice
Chair: Stuart Shaw, Room: Castelo 4-5

13:45 - 14:15 Equity in education within the European Union; A study based on PISA 2015 data
Remco Feskens[1, 2], Cor Sluijter[1]
[1]Cito, Netherlands
[2]Twente University, Netherlands

14:15 - 14:45 “Fair assessment” in a time of high-stakes testing and increasing student diversity in schools: The voice of three Chilean schools from a social justice perspective
Tamara Rozas[1, 2]
[1]University College London, United Kingdom
[2]Universidad de Chile, Chile

14:45 - 15:15 Assessment of students with special needs. Challenges, dilemmas and tensions between national regulations and teacher practices
Astrid Gillespie[1]
[1]Oslo Metropolitan University, Norway

Session AA: Papers 15-17 – Assessment of Practical Skills
Chair: Stéphanie Berger, Room: Castelo 10

13:45 - 14:15 Evaluating the impact of qualification reform on students’ practical skills
Stuart Cadwallader[1]
[1]Ofqual, United Kingdom

14:15 - 14:45 Re-heated meals: Revisiting the teaching, learning and assessment of practical cookery in schools
Gill Elliott[1], Jo Ireland[1]
[1]Cambridge Assessment, United Kingdom

14:45 - 15:15 Which factors play a role in explaining results on cognitive and practical skills in technology by 14- to 15-year-old students?
Lien Willem[1], Jan Ardies[2], Jaan Harnisfeger[1], Rianne Janssen[1]
[1]KU Leuven, Belgium
[2]Artesis Plantijn Hogeschool Antwerpen, Belgium

Session AAA: Papers 18-20 – Reliability
Chair: Anabela Serrão, Room: Castelo 3

13:45 - 14:15 A new standard setting procedure for competency based performances
Bas Hemker[1]
[1]Cito, Netherlands

14:15 - 14:45 Maximising the reliability and role of expert judgement in standard maintaining to account for changes in student performance
Milja Curcin[1], Beth Black[1]
[1]Ofqual, United Kingdom

14:45 - 15:15 Applying different measurement theories to evaluate marker reliability in vocational assessments
Zeeshan Rahman[1]
[1]City & Guilds, United Kingdom

15:15 - 15:45 Coffee break

Session F: Papers 21-23 – Formative Assessment
Chair: Karen Dunn, Room: Castelo 1-2

15:45 - 16:15 Practice of formative assessment in teacher education: case studies of Australia and Vietnam
Anh Duong[1]
[1]The University of Sydney, Australia

16:15 - 16:45 Student perspectives on formative feedback as part of writing portfolios in higher education
Sarah Hoem Iversen[1], Zoltan Varga[1], Monika Bader[1], Tony Burner[2]
[1]Western Norway University of Applied Sciences, Norway
[2]University of South-Eastern Norway, Norway

16:45 - 17:15 How summative assessments can be formative: using reading comprehension item data to inform teaching
Jemma Coulton[1], Anne Kispal[1]
[1]National Foundation for Educational Research, United Kingdom

Session G: Papers 24-26 – Assessing Receptive Skills
Chair: Andrew Boyle, Room: Castelo 9

15:45 - 16:15 Reading and Test-Taking Strategies used in the TOEFL Junior Standard Reading Test: Evidence from Retrospective Think-Aloud Protocols
Dina Tsagari[1], Trisevgeni Liontou[2]
[1]Oslo Metropolitan University, Norway
[2]Department of English Language & Literature / National & Kapodistrian University of Athens, Greece

16:15 - 16:45 Transformation during assessment: practice effects during the ESLC listening test
Andrés Christiansen[1], Rianne Janssen[1]
[1]KU Leuven, Belgium

16:45 - 17:15 Automated Scoring of Open-Ended Items for Reading Literacy Assessment in the Russian language
Maxim Skryabin[1], Alina Ivanova[1]
[1]National Research University Higher School of Economics, Russia

Session H: Papers 27-29 – National Tests and Examinations I
Chair: Bas Hemker, Room: Castelo 8

15:45 - 16:15 Post 16 Qualification Reforms: Impacts on mathematics teaching, learning and assessment in England
Ben Redmond[1], Jennie Golding[2], Grace Grima[1]
[1]Pearson UK, United Kingdom
[2]Institute of Education, United Kingdom

16:15 - 16:45 Qualifications Reform in Wales: Opportunities and challenges for high-stakes assessment
Oliver Stacey[1]
[1]Qualifications Wales, United Kingdom

16:45 - 17:15 The Yellow Wallpaper effect: The difficulties of moderating coursework
Stephen Holmes[1], Ellie Keys[1], Beth Black[1]
[1]Ofqual, United Kingdom

Session I: Papers 30-32 – Comparative Judgement I
Chair: Mary Richardson, Room: Castelo 6-7

15:45 - 16:15 A framework for describing comparability between alternative assessments
Stuart Shaw[1], Victoria Crisp[1], Sarah Hughes[1]
[1]Cambridge Assessment, United Kingdom

16:15 - 16:45 Why a Unified Approach to Language Scales Matters: The Case for Comparative Judgement
Rose Clesham[1], Sarah Hughes[1]
[1]Pearson UK, United Kingdom

16:45 - 17:15 Transforming the marking of extended responses through an understanding of complex judgment processes
Ayesha Ahmed[1]
[1]University of Cambridge, United Kingdom

Session J: Papers 33-35 – On-Screen Assessment
Chair: Lenka Fiřtová, Room: Castelo 4-5

15:45 - 16:15 Student engagement with on-screen assessments: A systematic literature review
Carla Pastorino[1]
[1]Cambridge Assessment, United Kingdom

16:15 - 16:45 On-screen assessments for young learners: Considerations for on-screen item type design and usage
Sanjay Mistry[1]
[1]Cambridge Assessment International Education, United Kingdom

16:45 - 17:15 The use of touchscreen vs. standard devices for marking high-stakes exams
Sarah Hughes[1], Martina Kuvalja[1]
[1]Cambridge Assessment, United Kingdom

Session BB: Papers 36-38 – Test Development II
Chair: Marieke van Onna, Room: Castelo 10

15:45 - 16:15 Developing command terms for assessment of performance and creating in the performing arts
Rebecca Hamer[1], Christina Haaf[1]
[1]International Baccalaureate, Netherlands

16:15 - 16:45 Tests as texts: investigating test questions from a sociolinguistic perspective
Filio Constantinou[1]
[1]Cambridge Assessment, University of Cambridge, United Kingdom

16:45 - 17:15 Quality criteria for assessment design
Paul Newton[1]
[1]Ofqual, United Kingdom

Session BBB: Papers 39-41 – Supporting Students' Performance
Chair: Gerry Shiel, Room: Castelo 3

15:45 - 16:15 Issues in using low-stakes assessment tools to identify and support at-risk students
Guri A. Nortvedt[1], Andreas Pettersen[1], Anubha Rohatgi[1], Karianne Berg Bratting[1]
[1]University of Oslo, Norway

16:15 - 16:45 Life gets in the way: Resits for students unable to present for high-stakes exams
Damian Murchan[1], Martyn Ware[2], Robert Quinn[2], Fabienne van der Kleij[3]
[1]Trinity College Dublin, Ireland
[2]Scottish Qualifications Authority, United Kingdom
[3]Australian Catholic University, Australia

16:45 - 17:15 Recommending learning materials to resit exam candidates using collaborative filtering
Eva de Schipper[1, 2], Remco Feskens[1, 2], Jos Keuning[1], Bernard Veldkamp[2]
[1]Cito, Netherlands
[2]Twente University, Netherlands

18.30 - 20.30 Events for members holding accreditation and for doctoral students
Location: Lisbon Geographical Society

Friday, 15th November

Session K: Papers 42-44 – International Surveys I
Chair: Beth Black, Room: Castelo 1-2

9:00 - 9:30 Transforming marking practice: the case of TIMSS 2019
Grace Grima[1], Mary Richardson[2], Tina Isaacs[2]
[1]Pearson UK, United Kingdom
[2]UCL Institute of Education, United Kingdom

9:30 - 10:00 The Test-Taking Behaviour of Irish Students in PISA 2015 and 2018: student engagement, interest, and concentration in computer-based assessment
Caroline McKeown[1], Sylvia Denner[1]
[1]Educational Research Centre, Ireland

10:00 - 10:30 Profiles of student motivation variables in grade-four TIMSS mathematics
Michalis Michaelides[1], Gavin Brown[2, 3], Hanna Eklöf[3], Elena Papanastasiou[4], Militsa Ivanova[1], Anastasios Markitsis[1]
[1]University of Cyprus, Cyprus
[2]University of Auckland, New Zealand
[3]Umeå University, Sweden
[4]University of Nicosia, Cyprus

Session L: Papers 45-46 – Assessing Mathematics I
Chair: George MacBride, Room: Castelo 9

9:00 - 9:30 MathemaTIC's Mini-Summative Assessments: Transforming Assessments in Digital Learning Pathways for Mathematics
Frauke Kesting[1], Carole Frieseisen[2], Filipe Lima da Cunha[2], Jacob Pucar[3]
[1]MENJE Luxembourg, Luxembourg
[2]SCRIPT - MENJE, Luxembourg
[3]Vretta Inc., Canada

9:30 - 10:00 MathemaTIC – using digital assessment to inform teachers of how students construct meaning in problem-solving
Amina Afif[1], Franck Salles[2]
[1]SCRIPT Data Division - Ministry of National Education, Luxembourg
[2]DEPP - Office of Student Assessment, Ministry of National Education, France

Session M: Papers 47-49 – Students' "Voice" in Assessment
Chair: Sarah Hughes, Room: Castelo 8

9:00 - 9:30 Students as stakeholders in the development of new assessment systems: A case study from Scotland
Martyn Ware[1], Shakeh Manassian[1]
[1]Scottish Qualifications Authority, United Kingdom

9:30 - 10:00 SQA Mental Health and Wellbeing Awards: meeting societal need with flexible learning and assessment
Jen Morrison[1], Elaine McFadyen[1]
[1]SQA, United Kingdom

10:00 - 10:30 Tracking test motivation in low-stakes large-scale assessment: the case of the National Reference Test (NRT) in England
Ming Wei Lee[1]
[1]Ofqual, United Kingdom

Session N: Papers 50-52 – Statistical Approaches to Assessment
Chair: Nico Dieteren, Room: Castelo 6-7

9:00 - 9:30 Rurality and educational attainment in Northern Ireland: A multilevel analysis
Gemma Cherry[1]
[1]Queen's University Belfast, United Kingdom

9:30 - 10:00 Measuring and Correcting the 'Sawtooth Effect' in a First Award
Elena Mariani[1]
[1]Pearson, United Kingdom

10:00 - 10:30 Exploratory Factor Analysis of the 2018 British Columbia Student Learning Survey
Todd Milford[1], Victor Glickman[1], John Anderson[1]
[1]University of Victoria, Canada

Session O: Papers 53-55 – E-Assessment
Chair: Yaw Bimpeh, Room: Castelo 4-5

9:00 - 9:30 Modeling the construct in a computerized performance-based assessment of ICT literacy
Georgy Vasin[1], Svetlana Avdeeva[1]
[1]Higher School of Economics, Institute of Education, Russia

9:30 - 10:00 The INVALSI computer-based assessment: psychometric challenges and opportunities in test design and score reporting
Marta Desimoni[1], Donatella Papa[1], Cristina Lasorsa[1], Rosalba Ceravolo[1], Antonella Costanzo[1], Angela Verschoor[2]
[1]INVALSI, Italy
[2]Cito, Netherlands

10:00 - 10:30 Transforming national examinations from paper and pen to an online mode of delivery: how easy is it? Egypt 2019 - a case study
David McVeigh[1]
[1]Pearson Qualification Services, United Kingdom

Session CC: Papers 56-58 – Assessment of Hard to Measure Skills
Chair: Tom Bramley, Room: Castelo 10

9:00 - 9:30 Measuring critical thinking through innovative assessment: An investigation of the dimensionality
Irana Uglanova[1]
[1]National Research University Higher School of Economics, Russia

9:30 - 10:00 Assessment of problem-solving skills
Martina Kuvalja[1], Stuart Shaw[1], Sarah Mattey[1], Giota Petkaki[1]
[1]Cambridge Assessment, United Kingdom

10:00 - 10:30 Is search for explicitly stated information really a lower-order cognitive skill in reading comprehension assessments?
Inna Antipkina[1], Ekatertina Aleksandrova[1], Alina Ivanova[2]
[1]Higher School of Economics, Russia
[2]National Research University Higher School of Economics, Russia

Session CCC: Papers 59-61 – Perceptions of GCSE
Chair: Thierry Rocher, Room: Castelo 3

9:00 - 9:30 Teacher interpretations of GCSE specifications: transformational knowledge in the classroom - studying a novel
Jenny Smith[1, 2]
[1]Independent researcher, United Kingdom
[2]University of Hertfordshire, United Kingdom

9:30 - 10:00 Teacher perceptions and experiences of the Non-Examination Assessment component of GCSEs in Wales: An exploration of fairness within the context of curriculum reform
Rachael Sperring[1], Kerry Jones[1]
[1]Qualifications Wales, United Kingdom

10:00 - 10:30 Unpacking the difficulty of GCSE Modern Foreign Language questions by combining subject expert ratings and objective item features
Tim Stratton[1], Nadir Zanini[1]
[1]Ofqual, United Kingdom

10.30 - 11.00 Coffee break

Discussion group 1, Room: Castelo 1-2

11:00 - 12:00 Reforming national examination systems: Assessing new competences emphasizing interdisciplinary learning, student collaboration and creativity
Sverre Tveit[1], Christian Lundahl[2]
[1]University of Agder, Norway
[2]Örebro University, Sweden

Discussion group 2, Room: Castelo 9

11:00 - 12:00 Transforming Assessment: How can digitalisation of high-stakes assessment enhance social inclusion through improved educational outcomes?
Irene Custodio[1], Kevin Mason[1], Grace Grima[1], Ellen Barrow[2]
[1]Pearson, United Kingdom
[2]Pearson Education, United Kingdom

Discussion group 3, Room: Castelo 8

11:00 - 12:00 Assess@Learning - digital formative assessment in classrooms: A European Project
Jannette Elwood[1], Kay Livingston[2], Patricia Wastiau[3]
[1]Queen's University Belfast, United Kingdom
[2]University of Glasgow, United Kingdom
[3]European Schoolnet Partnership, Belgium

12.00 - 1300 General assembly
Chair: Jannette Elwood, Room: Castelo 1-2

13.00 - 14.00 Lunch

Session P: Papers 62-64 – Language Issues in Assessment
Chair: Isabel Nisbet, Room: Castelo 1-2

14:00 - 14:30 Secondary school foreign language qualifications in England through the lens of the Common European Framework of Reference for Languages (CEFR): are assessment standards too high?
Milja Curcin[1], Beth Black[1]
[1]Ofqual, United Kingdom

14:30 - 15:00 Modern languages qualifications in Northern Ireland: student and teacher perceptions of difficulty, grading and decision-making
Leanne Henderson[1], Janice Carruthers[1], Ian Collen[1]
[1]Queen's University Belfast, United Kingdom

15:00 - 15:30 The CEFR as an assessment tool for learner linguistic and content competence: assisting learners in understanding the language proficiency needed for specific content goals in the CLIL classroom
Stuart Shaw[1]
[1]Cambridge Assessment, United Kingdom

Session Q: Papers 65-67 – Psychometrics II
Chair: Tim Oates, Room: Castelo 9

14:00 - 14:30 Equating by pairwise comparisons
Marieke Van Onna[1], Tecla Lampe[1]
[1]Cito, Netherlands

14:30 - 15:00 Balancing between psychometric validity and content validity: the case of differential item functioning for gender in a national assessment of French as a foreign language
Koen Aesaert[1], Jo Denis[1], Karen Van Renterghem[1]
[1]KU Leuven, Belgium

15:00 - 15:30 How can we use Item Response Times in the Low-Stakes Testing? Ideas on Reliability, Cross-National Comparability, and Responses Classification
Denis Federiakin[1]
[1]NRU Higher School of Economics, Russia

Session R: Papers 68-70 – School Improvement
Chair: Angela Verschoor, Room: Castelo 8

14:00 - 14:30 Developing a framework for school level data driven decision making to improve student achievements
Pāvels Pestovs[1], Dace Namsone[1], Ilze Saleniece[1]
[1]University of Latvia, Latvia

14:30 - 15:00 Performance evaluation in Nazarbayev Intellectual Schools: evidence from school inspections
Raigul Kakabayeva[1], Gulmira Zhailauova[1], Gulnar Kurmanbayeva[1], Olga Mozhayeva[1]
[1]Autonomous educational organization Nazarbayev Intellectual Schools, Kazakhstan

15:00 - 15:30 Preparing for high-stakes assessment of aspirational curricula: the role of educative resources
Alistair Hooper[1], Jennie Golding[2], Grace Grima[3]
[1]Pearson, United Kingdom
[2]University College London Institute of Education, United Kingdom
[3]Pearson UK, United Kingdom

Session S: Papers 71-72a – International Surveys II
Chair: Anton Béguin, Room: Castelo 6-7

14:00 - 14:30 Learning for or learning from PISA? Developing a tailor-made training course for sustainable transformation of education and assessment towards 21st century functional literacy
Nico Dieteren[1]
[1]Cito, Netherlands

14:30 - 15:00 Language effects in PIRLS 2016: Towards a more thorough analysis of differential item functioning
Yasmine El Masri[1], Joshua McGrane[1]
[1]University of Oxford, United Kingdom

15:00 - 15:30 Transforming Teaching, Learning and Assessment though TIMSS
Elena Papanastasiou[1], Maria Evagorou[1]
[1]University of Nicosia, Cyprus

Session Y: Papers 73-75 – Validity and Validation
Chair: Caroline Jongkamp, Room: Castelo 4-5

14:00 - 14:30 Assembled Validity: The Case of ILSAs
Bryan Maddox[1, 2], Bruno D. Zumbo[3], Camilla Addey[4]
[1]Assessment MicroAnalytics, United Kingdom
[2]University of East Anglia, United Kingdom
[3]University of British Columbia, Canada
[4]GEPS, Universitat Autònoma de Barcelona, Spain

14:30 - 15:00 Validation of the student selection system used for Nazarbayev Intellectual Schools
Aigul Jandarova[1], Zamira Rakhymbayeva[1], Aidana Shilibekova[1], Olga Mozhayeva[1]
[1]AEO Nazarbayev Intellectual Schools, Kazakhstan

15:00 - 15:30 Enhancing assessment validity through the use of animated videos: An experimental study comparing text-based and animated situational judgement tests
Anastasios Karakolidis[1], Michael O'Leary[2], Darina Scully[2]
[1]Centre for Assessment Research Policy and Practice in Education (CARPE), Ireland
[2]Dublin City University, Ireland

Session DD: Papers 76-78 – Comparative Judgement II
Chair: Saskia Wools, Room: Castelo 10

14:00 - 14:30 A Comparative Judgement Approach to the Large-Scale Assessment of Primary Writing in England
Christopher Wheadon[1], Daisy Christodoulou[1], Patrick Barmby[1]
[1]No More Marking Ltd., United Kingdom

14:30 - 15:00 Moderation of non-exam assessments: a novel approach using comparative judgement
Lucy Chambers[1], Sylvia Vitello[1], Carmen Vidal Rodeiro[1]
[1]Cambridge Assessment, United Kingdom

15:00 - 15:30 Judges' considerations in assessing children's writing in a comparative judgement process
Patrick Barmby[1], Daisy Christodoulou[1], Christopher Wheadon[1]
[1]No More Marking Ltd., United Kingdom

Session DDD: Papers 79-81 – Educational Approaches to Assessment
Chair: Ayesha Ahmed, Room: Castelo 3

14:00 - 14:30 Adapting the Cognitive Abilities Test (CAT4) to support teaching and learning in Chinese classrooms
Bernadetta Brzyska[1]
[1]GL Education, United Kingdom

14:30 - 15:00 What do student skills assessments tell us about performance gaps by gender?
Marianne Fabre[1], Lea Chabanon[1], Thomas Portelli-Tronville[1]
[1]Direction de l'évaluation, de la prospective et de la performance [DEPP], France

15:00 - 15:30 Two centuries of 'cram' – a history of cramming in UK educational assessment
Lydia May Townsend[1]
[1]Institute of Education, University College London, United Kingdom

15.30 - 16.00 Coffee break

16.00 - 18.10 Ignite Session and
16.00 - 17.00 Symposia

Ignite Session
Chair: Andrej Novik, Room: Castelo 1-2

16:00 - 16:10 Guidelines for formative tests in the classroom based on memory research
Desirée Joosten - ten Brinke[1, 2], Kim Dirkx[1], Gino Camp[1]
[1]Open University of the Netherlands, Netherlands
[2]Fontys University of Applied Sciences, Netherlands

16:10 - 16:20 How can we tell if it is valid? Using operational data to build an argument for validity
Kevin Mason[1]
[1]Pearson, United Kingdom

16:20 - 16:30 Reflex: A generic app for evaluation and monitoring of formative assessments
Hendrik Straat[1], Romy Noordhof[1]
[1]Cito, Netherlands

16:30 - 16:40 Getting out of their heads – using concept maps to elicit teachers' assessment literacy
Martin Johnson[1], Victoria Coleman[1]
[1]Cambridge Assessment, United Kingdom

16:40 - 16:50 Accelerating Innovations in Technology-Based Assessment
Mark Molenaar[1]
[1]Open Assessment Technologies, Luxembourg

16:50 - 17:00 Micro-Analysis in Large-Scale Assessment
Bryan Maddox[1, 2]
[1]Assessment MicroAnalytics, United Kingdom
[2]University of East Anglia, United Kingdom

17:00 - 17:10 Using PISA process data for evaluating the validity of self-reported test-taking effort and the impact of low effort on item performance
Hanna Eklöf[1], Peter Fjällström[1]
[1]Umeå University, Sweden

17:10 - 17:20 Enhancing Learning and Assessment Systems via Continuous Tracking of Practice Assessment Analytics & Personalized Resource Recommendations
Alina von Davier[1]
[1]ACT, United States

17:20 - 17:30 Do classroom assessment scores affect future academic outcomes?
Jennifer Vinas-Forcade[1, 2], Cindy Mels[3], Martin Valcke[1], Ilse Derluyn[1]
[1]Ghent University, Belgium
[2]Instituto Nacional de Evaluación Educativa, Uruguay
[3]Universidad Católica del Uruguay, Uruguay

17:30 - 17:40 What happens when assessments are digitalised?
Anna Lind Pantzare[1]
[1]Umeå University, Sweden

17:40 - 17:50 Combining proficiency measurement and mastery evaluation
Anton Béguin[1], Hendrik Straat[1]
[1]Cito, Netherlands

17:50 - 18:00 The importance of establishing the validity of assessments in educational experiments
Andrew Boyle[1]
[1]AlphaPlus Consultancy, United Kingdom

18:00 - 18:10 Redefining Student Success
Tanya Kolosova[1]
[1]YieldWise Inc, United States

Symposia

Room: Castelo 9
16.00 - 17.00 The rare but persistent problem of errors in examination papers and other assessment instruments
Irenka Suto, Paul Newton, Joanna Williamson, Sylvia Vitello, Nicky Rushton

Room: Castelo 8
16.00 - 17.00 Large Scale Digital Examns in Dutch Intermediate Vocational Education: Lessons Learned
Marcel Claessens, Peter Hakvoort, Maaike Beuving, Marieke van Onna, Rolf Vegar Olsen, Cor Sluijter

Room: Castelo 6-7
16.00 - 17.00 Developing, Analysing and Using: The Experience of the Scottish National Standardised Assessments and their Focus on Supporting Teachers
Sarah Richardson, Helen Claydon, Bethany Davies, Sladana Krstic, Anaghaa Wagh

Room: Castelo 4-5
16.00 - 17.00 Progression: Everyone is a Learner
George MacBride, David Morrison-Love, Jannette Elwood, Louise Hayward, Ernest Spencer, Kara Makara, Janine Barnes, Elaine Sharpling, Alex Southern, David Stacey, Jane Waters

19.00 - 23.00 Conference dinner
Location: Montes Claros Restaurant

Saturday, 16th November

Session T: Papers 82-84 – Assessing Mathematics II
Chair: Grace Grima, Room: Castelo 1-2

9:00 - 9:30 Assessing mathematics competence in Design and Technology: policy intentions and practical outcomes
Cesare Aloisi[1], Gemma O'Brien[1], Sarah Carter[1], Stephen Wooding[1]
[1]AQA, United Kingdom

9:30 - 10:00 The Nordic student experience: How do students in Finland, Norway and Sweden experience instructional quality in Language Arts and Mathematics?
Astrid Roe[1], Marte Blikstad-Balas[1], Michael Tengberg[2]
[1]University of Oslo, Norway
[2]Karlstad University, Sweden

10:00 - 10:30 Investigation of Heterogeneity in Mathematics Abilities across Compulsory School through Vertical Scaling
Stéphanie Berger[1], Laura Helbling[1], Martin J. Tomasik[1, 2], Urs Moser[1]
[1]University of Zurich, Switzerland
[2]University of Witten/Herdecke, Germany

Session U: Papers 85-87 – Psychometrics III
Chair: Guri A. Nortvedt, Room: Castelo 9

9:00 - 9:30 Looking beyond the test scores: Latent motivational profiling of teenage English language learners from four country contexts
Karen Dunn[1]
[1]British Council, United Kingdom

9:30 - 10:00 Screening System for Professional Training Programs for Israeli School Principals: Development, Operation and Validation
Avital Moshinsky[1], David Ziegler[1], Lisa Levy[2], Revital Nachum[2], Itay Soudry[2], Anat Shirazi[3], Helena Kimron[2], Hani Shilton[4]
[1]NITE, Israel
[2]Avney Rosha Institute, Israel
[3]Ministry of Education, Israel
[4]The Open University, Israel

10:00 - 10:30 Validity and Validation of Formative Assessment
Saskia Wools[1]
[1]Cito, Netherlands

Session V: Papers 88-90 – International Surveys III
Chair: Jean-Pierre Jeantheau, Room: Castelo 8

9:00 - 9:30 What do international large-scale assessments tell us about high achievement in mathematics and science, with specific reference to Ireland and some comparison countries?
Vasiliki Pitsia[1], Michael O'Leary[1], Gerry Shiel[2], Zita Lysaght[1]
[1]Dublin City University, Ireland
[2]Educational Research Centre, Dublin, Ireland

9:30 - 10:00 Relationships between 15-year olds' access to technology, perceived competence, autonomy and attitudes to ICTs, and their performance on PISA 2015 science in Ireland
Sarah Mc Ateer[1], Lynsey O'Keeffe[1], Gerry Shiel[1], Caroline McKeown[1]
[1]Educational Research Centre, Dublin, Ireland

10:00 - 10:30 The ability to read numbers: A universal measure?
Elena Kardanova[1], Dmitrii Kholiavin[1], Peter Tymms[2], Christine Merrell[2]
[1]National Research University Higher School of Economics, Russia
[2]Durham University, United Kingdom

Session W: Papers 91-92 – Policy
Chair: Paul Newton, Room: Castelo 6-7

9:00 - 9:30 The 'grey history' of assessment: understanding the origins of England's new model of assessment of practical work in Science
Tim Oates[1]
[1]Cambridge Assessment, United Kingdom

9:30 - 10:00 Irish Primary Teachers' Use of and Attitudes to Standardised Achievement Testing in English Reading and Mathematics
Zita Lysaght[1, 2], Deirbhile Nic Craith[3], Michael O'Leary[1, 2]
[1]Dublin City University, Ireland
[2]Centre for Assessment Research Policy and Practice in Education (CARPE), Ireland
[3]Irish National Teachers' Organisation, Ireland

Session X: Papers 93-95 – Assessment and Teachers´ Practice
Chair: Roger Murphy, Room: Castelo 4-5

9:00 - 9:30 Corpus-based teaching practices & classroom-based assessment: Putting theory into practice
Trisevgeni Liontou[1]
[1]Department of English Language & Literature / National & Kapodistrian University of Athens, Greece

9:30 - 10:00 Assessment Literacy – How does being an examiner enhance teachers' understanding of assessment?
Victoria Coleman[1], Martin Johnson[1]
[1]Cambridge Assessment, United Kingdom

10:00 - 10:30 Understanding educators' classroom assessment needs: Using human-centered design principles to establish an assessment use case
Leanne Ketterlin Geller[1], Tina Barton[1], Lindsey Perry[1]
[1]Southern Methodist University, United States

Session EE: Papers 96-98 – National Tests and Examinations II
Chair: Rose Clesham, Room: Castelo 10

9:00 - 9:30 Age-standardising on-demand tests: Is there an effect of "learning time"?
Ben Smith[1]
[1]AlphaPlus, United Kingdom

9:30 - 10:00 Exploiting the longitudinal data from exhaustive assessments to measure skills and progress during the first years of schooling
Marianne Fabre[1], Thomas Portelli-Tronville[1], Léa Chabanon[1]
[1]Direction de l'évaluation, de la prospective et de la performance [DEPP], France

10:00 - 10:30 Predicting grades in external summative assessment of graduates: example from Nazarbayev Intellectual Schools
Daulet Shadiyev[1], Zamira Rakhymbayeva[1], Yerbol Almenov[1]
[1]Nazarbayev Intellectual Schools, Kazakhstan

Session EEE: Papers 99-101 – National Tests and Educational Change
Chair: Sandra Johnson, Room: Castelo 3

9:00 - 9:30 External evaluation as a tool for school development: how do Flemish teachers and school leaders engage with school-level feedback from large-scale national assessments?
Evelyn Goffin[1, 2], Mieke Heyvaert[2], Isabel Laenen[2], Rianne Janssen[2], Jan Vanhoof[1]
[1]University of Antwerp, Belgium
[2]KU Leuven, Belgium

9:30 - 10:00 Census evaluation as a tool to support educational change: the case of Science education in Peru
Yoni Arámbulo Mogollón[1], Carmen Maribel Carpio[2], Caroline Jongkamp[3]
[1]UMC, Oficina de Medición de la Calidad de los Aprendizajes, Peru
[2]UCR, Universidad Nacional de Costa Rica, Costa Rica
[3]Cito, Institute for Educational Measurement, Netherlands

10:00 - 10:30 Implementation of a Pupil Monitoring System on Curaçao to enhance learning outcomes
Wil Knappers[1], Esther Padilla-Bomberg[2], Frans Kleintjes[1], Servaas Frissen[1]
[1]Cito, Netherlands
[2]Curaçao Expertise center for Tests & Exams, Netherlands Antilles

10.30 - 11.00 Coffee break

11.00 - 11.45 Keynote speech
Chair: Rolf V. Olsen, Room: Castelo 1-2
Title: Improving student's performance with Active Learning
Prof. Xavier Gimenéz

11.50 - 12.35 Keynote speech
Chair: Christina Wikström, Room: Castelo 1-2
Title: Using Response Process Data for informing Group Comparisons
Prof. Kadriye Ercikan

12.35 - 13.00 Awards and Closing Session
Chair: Jannette Elwood, Room: Castelo 1-2

13.00 - 14.00 Lunch

Thursday, 14th November

9.30 - 10.15 **Keynote speech**
Chair: Jannette Elwood, Room: Castelo 1-2
Title: Towards a New Generation of External Assessments: Reflection on a Systematic Literature Review
Prof. Domingos Fernandes

Short bio
Domingos Fernandes is a full and tenured professor of Educational Evaluation and also an integrated researcher of the Research and Development Unit in Education and Training at the Institute of Education of the University of Lisbon. Currently, he is serving as coordinator of the Department of Education and Training Policies and of the Master's and Doctoral programs in Educational Evaluation as well. His main research and teaching interests are Evaluation Theory, Program Evaluation, Policies Evaluation, and Learning Assessment. He has been a visiting professor in a number of foreign universities such as Texas A&M University in the USA, University of São Paulo (USP) and State University of São Paulo (UNESP) in Brasil, and University of La Salle in Colombia. Moreover, he has been the principal researcher and coordinator of several financed national and international research and evaluation projects. He is the author of more than one hundred publications (e.g., research journal articles, books, book chapters, monographs, research and evaluation reports).

10.15 - 10.45 Coffee break

10.45 - 11.30 **Keynote speech**
Chair: Alex Scharaschkin, Room: Castelo 1-2
Formative Assessment in Student Transition to Higher Education - A Sociocultural Perspective
Dr. Aisling Keane, Kathleen Tattersall, New Assessment Reseacher

Abstract
Informed by Rogoff's Three Planes of Analysis framework and influenced by situated participationism the Kathleen Tattersall New Researcher Award keynote lecture will explore and expand discussions surrounding formative assessment to offer alternative approaches to current practices dominant in the early years of undergraduate teaching and educational transitions in general. This is important as a clearly articulated sociocultural perspective provides comprehensive theoretical insight into formative assessment practices in first year higher education which inadvertently negatively impact on student enculturation into a new community. This work advocates for sociocultural approaches which see pedagogies as transformative for newcomers when they rely on clear frameworks of mutual participation between staff and students in valuable on-going cultural activities. Such pedagogies facilitate learner involvement to recognise processes and efforts which contribute to community goals.

Short bio
Upon completing her Ph.D in Anatomy (National University of Ireland, Galway, Ireland) Aisling joined the Centre for Biomedical Sciences Education at the Queen's University Belfast (QUB) Northern Ireland in 2005 as a Lecturer (Education). Recognising the importance of educational research in third level education Aisling undertook and graduated with a Doctorate in Education (2019) from QUB. Her educational research is underpinned by sociocultural approaches to exploring the nature of assessment, learning and student transition to third level education and scholarship of teaching and learning in Higher

Education. Aisling's work makes an original contribution to the field through the application of a sociocultural framework to explore student experiences of formative assessment in the first year of university and the impact of this on subsequent approaches to assessment and learning, particularly in the second year.

Poster Presentations

11.30 - 12.45 Posters
Chair: Cor Sluijter, Room: Castelo 1-2

Poster 1 How are Abacus resources supporting transformational mathematics learning and assessment?

Ellen Barrow[1], Jennie Golding[2]
[1]Pearson Education, United Kingdom
[2]University College London Institute of Education, United Kingdom

The English primary mathematics curriculum was reformed from 2014, reflecting aspirations for transformational mathematical functioning for the twenty-first century. We report on findings from a longitudinal study exploring the impact of a related and widely-used set of curriculum and assessment materials. We set out to explore the motivations for adoption, how the different elements are typically used in schools, and perceptions of their effectiveness in meeting the needs of teachers and learners, including perceived priorities for further development of the resources. In particular, we studied the use of the resources to support children's grasp of challenging curriculum areas. Qualitative data from observations, interviews and focus groups, supplemented by progression data, were collected from teachers and children in nine lower primary and nine upper primary classes, each class followed through two years' enactment, and from the schools' mathematics coordinators. We identify significant findings, such as the role ascribed to digital elements, the apparent impact of resource use on learner and teacher affect, and its sometimes transformational impact on children's engagement with key mathematical processes such as problem solving, reasoning and fluency. We outline the consequent actions taken to further develop efficacy.

Poster 2 System assessments for transformation: uniting forces of national assessments and the inspectorate in Flanders (Belgium)

Mieke Heyvaert[1], Ward Adelhof[2], Rianne Janssen[1]
[1]KU Leuven, Belgium
[2]Ghent University, Belgium

Both national assessments and the inspectorate monitor the quality of education at the system level. Whereas national assessments give a detailed account of student performances in a certain topic, the inspectorate audits schools on diverse aspects of the provided education. National assessments only have access to characteristics of schools, teachers, and students through background questionnaires, whereas the inspectorate carries out school visits in which they talk to representatives of all stakeholders of a school and even may observe classroom interactions. Given their complementary view on the quality of education, the Flemish policy research centre for test development and assessments and the Flemish inspectorate recently explored how they can structurally cooperate using a partially mixed concurrent equal status mixed methods design throughout the different research phases of a specific national assessment (i.e., exploration, test development, data collection and analysis, and valorisation). The aim of this cooperation is first to present a broader view of the quality of education in a certain topic to schools, teachers, governance agencies, educational researchers, and other educational stakeholders, and second to enhance actions taken by the educational field to further improve educational quality and, hence, the performance of future student cohorts.

Poster 3 Validity considerations in digital iterations of PISA: What can process data tell us about test-taking behaviour amongst students engaging with science assessments?

Caroline McKeown[1, 2]
[1]*Educational Research Centre, Ireland*
[2]*School of Education, Trinity College Dublin, Ireland*

Use of log-file and process data is a recent and evolving area of educational research, characterised by big data and considerable complexity in analysis and interpretation. This study explores student processes and test-taking behaviour captured in log-files, from the Programme of International Student Assessment (PISA), which transitioned to computer-based assessment in 2015. Structured methods were used to extract process data from student log-files drawn from PISA 2015 in Ireland, for the domain of scientific literacy. Analyses focused in particular on the new interactive science simulation items, in which students in Ireland underperformed when compared to non-interactive and trend items. Processing of these data has resulted in the development of meaningful sequences related to 21st century competencies, such as critical thinking and problem-solving. Findings from the study highlight the specific strategies employed by students when engaging with interactive and non-interactive items, and suggest possible reasons for the significant drop in the performance of Irish students on PISA science in 2015. A number of implications for validating inferences drawn from student responses to the new interactive science items in PISA are provided.

Poster 4 Examining the impact on student performance in Reading, Mathematics and Science in PISA, from the perspective of teachers and principals, when students are tested at different times of the year (autumn versus spring testing)

Sylvia Denner[1, 2]
[1]*Educational Research Centre, Ireland*
[2]*PhD Candidate, Dublin City University School of Policy and Practice, Ireland*

In Ireland, consideration is being given to changing the test window for PISA. This study examines the impact on performance in PISA when testing at different times of the year, and the factors associated with it. A representative sample of students participated in an autumn study after PISA 2018 Main Study data collection earlier in the spring. Teacher and principal views were collected via questionnaires and interviews after testing in schools.
This poster focuses on results from the teachers' and principals' perspectives in the areas of student engagement, motivation, time spent re-visiting topics at the beginning of an academic year and their reasons why a student might perform differently at different times of the year. Findings from the teachers' questionnaires point towards a differential impact of the change of testing for students, with a greater anticipated impact perceived for middle-achievers and less impact anticipated for high and low-achievers. Principals referenced the study design of PISA and the age-based sample as a possible factor in negating an impact on performance. The responses from teachers and principals provide a fuller understanding of the implications of a possible change in the timing of the administration of PISA in Ireland.

Poster 5 Using a comparative judgement method to assess inter-board equivalence of reformed GCSE (9-1) Mathematics question papers

Faiza Tufail[1], *David McVeigh*[2]
[1]*Pearson, United Kingdom*
[2]*Pearson Qualification Services, United Kingdom*

Any qualification reform will necessitate an amount of change. The reform of the GCSE (9-1) Mathematics qualifications in the UK which took place in 2017 was no different.

In the UK there are multiple providers (Awarding Organisations) each developing their product offer, using the curriculum content set by the UK Government's Department for Education (DfE), and the statutory qualification and subject level rules and guidance issued by the Office of Qualifications and Examinations Regulation (Ofqual).
If the qualifications fulfill the regulatory requirements they will be accredited by Ofqual. This may lead to qualifications that adhere to the regulatory framework, but with slightly different but allowable variations in the interpretation of the government-issued materials.
As this concept is of significant importance, our Assessment Design team administered a Comparative Judgement exercise using assessment materials across the three largest Awarding Organisations for the accredited GCSE (9-1) Mathematics qualifications Summer 2018 series.
This poster is of interest to personnel working in awarding bodies and all professionals involved in education and qualification reforms.

Poster 6 Developing CEFR-based Descriptors for the Assessment of Competency in Turkish as a Foreign Language

Yiğit Savuran[1]
[1]*Anadolu University, Turkey*

Turkish as a foreign language (TFL) has gained widespread attention over the last decade for various political and regional reasons. Turkish, which is one of the most significant languages to learn especially in different parts of Europe, Western Asia, and Middle East, is now attracting much more attention due to huge number of immigrants and asylum-seekers in Turkey. This growing interest in TFL has led many educational professionals and scientists to study various aspects of it. Many studies focusing on topics such as curriculum and planning, material development, teaching and learning methods and strategies have been carried out by researchers. However, the assessment and evaluation of TFL has not gained much popularity and remained considerably underdeveloped. Therefore, as a part of ongoing Ph.D. dissertation, this study focuses on developing CEFR-based descriptors for the assessment of main competency areas in language, namely written and oral reception, production and interaction. Taking various CEFR descriptors in different scales as well as many others from other resources (e.g. Pearson's GSE), the study relates different types of language examinations to CEFR levels and scales. By doing so, it also aims to set a standard and benchmarking for defining the competency level of a TFL learner.

Poster 7 From paper-based to computer-based assessment of numeracy: some consequences for item design

Karianne Berg Bratting[1], *Guri A. Nortvedt*[1], *Andreas Pettersen*[1], *Anubha Rohatgi*[1]
[1]*University of Oslo, Norway*

In January 2019, the third generation of the Norwegian mapping tests in numeracy for primary grade students was initiated. The purpose of the tests is mainly the same: to identify students at risk of lagging behind in mathematics while focusing on counting skills, number concepts, and computing skills. The first two generations of the test were paper-based while the new ones will be computer-based. This transition raises potential challenges as unintended visual support in the computer-based format could influence the difficulty of tasks adapted from the paper-based format. For instance, sorting numbers by size on a paper-based assessment demands students to carefully consider each number while holding them all in their mind (to identify the smallest number, for instance). In a digital format, where students can move numbers around, they might compare them side-by-side. Similarly, designing items in which students must order numbers by size is challenging as a keyboard can provide visual support, e.g. the order of numbers. The purpose of the poster is to display items in a paper-based and computer-based format to discuss the potential threats to validity that visual support might cause.

Poster 8 The Power of Data to take smart decisions for school improvement

Senad Karavdic[1], Amina Afif[1], Graziella Losciale[1]
[1]SCRIPT, Luxembourg

In the age of the accountability and the evidence-based decision making in education, the data literacy has become a priority in schools for their continuous improvement process. With six partners countries, the DATADRIVE project (ERASMUS +) aims to explore innovative ways of empowering schools in their data use.
At an early stage of a project, the literature review and qualitative data analysis performed on students, teachers and school leaders will allow us to create a data-driven school improvement framework. At the second stage of a project, questionnaire will be administered to school teachers in order to measure the extent of data use to drive school improvements.
These outputs will provide a content for three training sessions for Professional Learning Community (PLC) in each school. In addition, this work will gather evidence from participating countries to create corresponding teaching material, videos and a handbook on data-driven school improvement. At the end of the project, workshops with all stakeholders will be held in each participating country, to familiarize them with methods of collecting and analysing data and how project outputs can be used to improve the way decision-making is done in schools.

Poster 9 Assessment for transformation at the school level? The use of parallel tests in Flanders (Belgium)

Isabel Laenen[1], Evelyn Goffin[2]
[1]Catholic University of Leuven, Belgium
[2]KU Leuven, Belgium

Primary and secondary schools in Flanders (Belgium) can administer parallel tests to their students in order to acquire school-level feedback. Parallel tests are developed alongside tests used in a large-scale national assessment of a certain topic, on a comprehensive measurement scale. Feedback on these tests comprises output data (Are attainment targets met?) and benchmarks (How does the school's performance compare to the national average?). Schools can use this information for internal quality assurance.
The poster will firstly elaborate on the purpose of this low-stakes external assessment tool. From a school development perspective, we want to equip schools with information that allows them to increase the quality of their educational outcomes. Parallel test results can instigate schools to do further research and triangulate with other forms of assessment, and stimulate them to (keep) systematically monitor(ing) their performance.
Secondly, we will illustrate the instrument's design and how the results are presented. Caterpillar plots provide a relative comparison to sample schools and represent schools' added value.
Thirdly, we will highlight recent developments and challenges, such as the inclusion of individual pupil results (presented with confidence intervals), the issue of data literacy, and the involvement of other educational stakeholders.

Poster 10 In Search of Assessment that is Fit for Purpose in Character and Citizenship Education

Ng May Gay[1], Osman Abdullah[1]
[1]Ministry of Education, Singapore

In Singapore, Character and Citizenship Education (CCE) aims to develop the character, social emotional well-being and citizenship dispositions of our students from the primary to post-secondary levels (7-18 years old). As CCE mainly concerns the development of values, dispositions and attitudes, assessment in CCE poses a challenge to schools, as they grapple

with the desire to know if what they do for CCE has any impact on their students. This is especially because the development of measures to assess social emotional skills or character development is faced with numerous difficulties. Furthermore, the integration of the assessment of social emotional learning and character into a single measure has yet to be accomplished (Elias, et. al., 2016). With the scarcity of literature on approaches to assessment in CCE, this study seeks to present a case for self-referential, sustainable assessment to reframe the discourse on assessment to focus on student learning and the development of competencies to make complex judgements for continuous character growth. Through a multiple-case study, why, how and when teachers apply the principles of self-referential, sustainable assessment in the approach to assessment in CCE will be shared.

Poster 11 Effects with learning aids on mathematical performance in vocational education in Flanders (Belgium)

Margo Vandenbroeck[1], Lien Willem[1], Rianne Janssen[1]
[1]KU Leuven, Belgium

Given students' familiarity to use learning aids (e.g., formulary, conversion table, roadmap) during mathematical lessons and evaluation in vocational education, teachers argue that students should also be able to use learning aids during a forthcoming national assessment in mathematics in Flanders. In order to investigate the teachers' belief that students will perform worse without learning aids, two modes of test administration were created (with and without learning aids) to which schools were randomly assigned. Using IRT and multilevel analysis, data collected in May 2019 from about 4000 14-year-old Flemish students in vocational education will be analyzed. The comparison between the two groups is made possible through the use of anchor items in the knowledge test. We will control for background variables (e.g., mindset of teachers, confidence, motivation,...) collected through teacher and student questionnaires. Past research (e.g., De Ruyck & Desoete, 2010; Säljö, Eklund, & Mäkitalo, 2006) predict either an advantage in performance for students using learning aids (as these tools may compensate for lacking knowledge/skills and reduce the strain on their short-term memory) or no advantage (given that mathematical problem solving also depends on formulating a mathematical model and relying on the right reasoning processes and solving strategies).

Poster 12 The Effect of MathemaTIC's New Summative Assessment Format in Digital Learning Pathways for Mathematics

Arbana Miftari[1], Jacob Pucar[1]
[1]Vretta Inc., Canada

Over the past five years, the Luxembourg Ministry of Education, in partnership with national and international experts, has developed over 500 interactive assessment-for-learning items designed to make mathematics relatable, tangible, and engaging for students. This assessment-for-learning resource is called MathemaTIC.
MathemaTIC's personalized learning environment contains learning pathways that are separated into different modules, each of which covers a key topic area that students work towards mastering. Each of the learning pathways is comprised of three different phases with unique item types: Learn, Practice, and Apply. Although the learning pathways and subsequent summative assessments proved to effectively engage students in their math education and provide teachers with an in-depth view of their students' performance at the end of each module, MathemaTIC has now introduced an additional layer of assessment items which will enhance both student engagement and performance monitoring. This new layer of assessment is comprised of mini-summative assessments that are spread throughout the modules.
Through this poster presentation, we will be highlighting the addition of this new assessment format into the MathemaTIC platform as well as showing comparative data that depicts the effect that the addition of this new assessment format has had on student success on the MathemaTIC platform.

Poster 13 From opinion to evidence: transforming organisational culture in two Awarding Organisations

Alison Rodrigues[1], Sarah Hughes[1]
[1]Cambridge Assessment International Education, United Kingdom

Research evidence is just one factor influencing decision making. Other factors include fit with culture; practicality; financial considerations; colleagues' beliefs; dominance of personalities; professional wisdom and policy and practice. We believe that research evidence should be a key driver for decision making. To this end a research-use initiative was launched based on a monitoring and evaluation framework and applied in two Awarding Organisations. Dimensions of the framework are: strategy and direction (what is the vision and mission?), research management (what processes and protocols are in place?), outputs (what mechanisms for sharing research are in place and are they appropriate?), uptake (are people accessing and sharing research?), impact (has research had any longer term impact?) and context (how have other factors affected impact?). Research use in the two Awarding Organisations is compared in terms of the aspects of the framework. Differences in research-use across the two Awarding Organisations can be described in terms of contextual factors such as: organisational culture including leadership; regulatory pressures, level of embeddedness of the research function; whether research is expected to justify or inform decision making; how agile the research process is; intended audience; appetite for research evidence.

Poster 14 From check to act: Involving stakeholders in resonating national assessment results back to educational practice

Sabine Dierick[1], Rianne Janssen[1], Koen Aesaert[1]
[1]KU Leuven, Belgium

Despite their wealth of information, national assessments not always succeed to create a positive 'washback effect' to the educational system by enhancing the quality of education and improving the learning outcomes of future student cohorts in the tested field. Frequently, national assessments only form a 'check' of the quality of education, but not enough effort is given to formulating subsequent 'actions'. In the present study, a qualitative participative research design is proposed to involve educational stakeholders in resonating national assessment results back to educational practice. More specifically, a focus group discussion with relevant stakeholders leads to contextualized interpretations of the results of a national assessments. Subsequently, a large, open hearing is organized in which these interpretations and possible actions are discussed in two rounds of round-table discussion, in which a moderator ensures the application of the deliberative method. As a result, a report of possible actions is made that can be used by educational policy makers to further plan and implement future educational reforms, in which a future national assessment completes the full 'plan-do-check-act' cycle. The proposed approach is illustrated with the results of a national assessment of Dutch as a first language in primary education in Flanders (Belgium).

Poster 15 The influence of differentiation on the quality of teaching gifted children

Mukhammed Mussabekov[1], Saule Vildanova[1], Nina Kashavarova[1]
[1]Center for Pedagogical Measurements under the AEO "Nazarbayev Intellectual Schools", Kazakhstan

This study focuses on the impact of differentiation on the quality of education of gifted children. The main question of the study: how does the application of differentiation affect the dynamics of the development of gifted students, and contribute to improve the quality of knowledge?
The study was based on the analysis of data of monitoring the implementation and application of the special programme "Teaching gifted children at school" in twenty

Nazarbayev Intellectual Schools. The features of these schools are that students are accepted through competitive selection. Teaching students caused great difficulties for teachers, many of whom experienced difficulties in choosing technologies of teaching, and strategies of assessing gifted students. Therefore, a programme of teaching pedagogical staff the methods of working with gifted students was developed. It was important to determine how the application of differentiation contributed to satisfying the unique needs of children with identified giftedness. The subject of the research is the data collected during the monitoring process.
The results obtained during the study allowed to determine the prospects for the further development of teaching gifted children practices, as well as to develop specific recommendations on the implementation of differentiation in other schools of Kazakhstan.

Poster 16 Potential threats to validity through the use of extended response items

Gillian Mann[1]
[1]*Scottish Qualifications Authority, United Kingdom*

In 2017, SQA convened #SQAfutures Vision Panel as a representative voice of young people to consider the future of assessment. The Panel noted that extended writing was useful 'but it may not be appropriate in all subjects because it could disadvantage candidates who are not good at communicating this way' (#SQAfutures, 2018).
In response, SQA examined the validity of using extended response items, across National Courses, particularly in relation to the assessment and reward of cognitive and/or communicative competence.
Findings indicate that extended response items form a significant feature of National Course question papers but communicative competence is only rewarded as appropriate to the course aims and purposes in the Languages, Humanities and one Social Science. As such, candidates are not being unfairly disadvantaged through poor communicative ability within any question papers. However, communicative competence is assessed and rewarded across a significant number of National Courses as extended writing within coursework. Subsequently, there is a danger that a candidate's ability to gain marks for higher order skills may be constrained by poor levels of communication.
The purpose of this poster is to showcase why the use of extended response items should be controlled as a matter of policy.

Poster 17 Vocational learners' perceptions on making summative assessment engaging

Vasile Rotaru[1]
[1]*Qualifications Wales, United Kingdom*

The term 'learner engagement' refers to learners enjoying and being involved in their learning/assessment. Studies have revealed an interaction between test and item characteristics, on the one hand, and motivation, effort and performance, on the other. Engaging assessment improves exam results, impacts learning strategies and helps traditionally disadvantaged students.
However, improving student engagement while taking tests has not been the focus of the tests' designers (Bae and Kokka, 2016). Various factors, such as the difficulty to operationalise the concept, the issues of bias and accessibility, and relative 'paucity of research', might have contributed to this situation. Lack of relevant research is even more noticeable in vocational education.
This poster describes an exploratory study to investigate students' engagement in assessment within a vocational context. Fourteen learners following the same course in a further education college have completed a survey measuring their attitudes towards learning. Once they undertake their summative assessment, they will attend a focus group to share their views on how engaging they felt the assessment was. The focus groups will explore the following aspects:
a) How learners view/define engagement?
b) What contributes to learners (dis)engagement in assessment?
c) What reasonably can be done to (re-)engage learners in assessment tasks?

Poster 18 The role of motivation in performance on mathematics

Naomi Carpentier[1], Lien Willem[1]
[1]KU Leuven, Belgium

In the spring of 2018, a large scale assessment was conducted on the mathematical skills of 2985 fourteen year old Flemish students (Carpentier al., 2019). From this research, it appears that in comparison with a previous assessment of mathematics in 2009 (Gielen et al., 2010), self-reported motivation for studying mathematics has grown. However, results on mathematical tests have not consistently improved.

Exploratory analyses on these data have shown that this growth in motivation differs between boys and girls. While motivation has stayed more or less the same for boys, girls have taken a big leap in catching up.
This begs the question whether an interaction effect can be found between gender and motivation. Results show that this is the case for 3 out of 9 mathematical subtests (number sense, algebra and geometric conceptualization). On these subtests, the correlation between motivation and performance is significantly smaller for girls than for boys. These findings could partially explain why motivation rising has not had the expected positive influence on performance.

Poster 19 Higher Applications of Mathematics

Kevin Gibson[1], Martin Brown[1]
[1]Scottish Qualifications Authority, United Kingdom

Higher Applications of Mathematics
Across the world, policy makers are recognising the importance of developing a STEM literate society, one that can survive and thrive in the modern world. The current workforce need skills and competencies that were not needed or particularly valued in the past – probability and statistical literacy, applying critical thinking, and an ability to reason mathematically.
In Scotland, there is an existing Higher in mathematics which is already recognised as an excellent award. To offer an alternative and equivalent to Higher Mathematics, the Scottish Qualifications Authority (SQA) are currently developing a Higher Applications of Mathematics. It will equip learners with the skills needed to interpret and analyse numerical and graphical information, simplify and solve problems, assess risk and make informed decisions.
Find out more about our progress on this journey.

Poster 20 Open questions in chemistry and physics: a creative approach to assessments of depth of knowledge and understanding of the science

Shakeh Manassian[1]
[1]Scottish Qualifications Authority, United Kingdom

SQA National Course assessments in Physics and Chemistry have always drawn on a variety of item types (short answer, constructed response, MCQ) to elicit candidate responses to the varied aspects of the Course content being assessed. One such item type which has been innovative in the Scottish context is the open-ended questions. These questions aim to promote a deeper understanding of the subject in the classroom as well as elicit more complex, integrative approaches to problem solving in the context of the external assessment. They provide candidates with an opportunity to show their understanding of the concepts they have been studying and whether they can draw on these concepts to address problems that are new and novel. There is no one correct answer to the question being set, and the mark scheme allows markers to judge candidate responses in relation to their insightfulness as well as their creativity and analytical thinking.
The open-ended questions have been in use since 2010. The purpose of this research is to

evaluate how they have evolved. By sharing our findings we hope to engage with our colleagues and widen the discussion about the skills which we aim to develop in our learners.

Poster 21 The GCSE Mathematics Saga...

Vasu Krishnaswamy[1, 2], Jennie Golding[2]
[1]Pearson UK, United Kingdom
[2]University College London Institute of Education, United Kingdom

Prior to 1988 young people in England were typically tracked for secondary mathematics education – high achievers entered for 'O level Mathematics' and others for 'CSE Mathematics', which had a quite different focus. This was seen to be undesirable and reform arrived with the 'GCSE' for all in 1988.
31 years later, what does the qualification mean for young people, teachers and employers today? Does it achieve what is intended? Mathematics learning is an important feature of any country's education system. Does a grade 4 (pass) in GCSE Mathematics, for example, unequivocally assure the stakeholders of the acquisition of key valued mathematics skills? This presentation outlines the purpose of the current 'transformational' GCSE Mathematics and summarises recent studies of the qualification that evidence the extent to which that is being achieved. Analysis of student and teacher interviews and surveys, and examination performance data, reveals that some unintended consequences have arisen. Systemic obstacles to gaining effective mathematics skills are identified, and an exploration of alternative approaches to assessment is presented. Is it time to consider a transformation to the way we assess mathematics at this level, so as to support student outcomes more coherent with qualification intentions?

Poster 22 The Scandinavian legacy of resisting formal grading – paradoxes and dilemmas

Sverre Tveit[1], Lise Vikan Sandvik[2], Henning Fjørtoft[2]
[1]University of Agder, Norway
[2]NTNU Norwegian University of Science and Technology, Norway

Grading has been a key component of assessment systems for centuries. Recently, international research has given more attention to potentially harmful effects on student learning. Policymakers are therefore seeking to reform assessment practices based on the idea that assessment can support learning and instruction, and that improving assessment therefore can improve learning. The Scandinavian countries have a long-standing tradition of prohibiting formal grading in primary education. This paper reports on a qualitative content analysis (QCA) of policy documents underpinning educational assessment policies in Norway from the 1939 to 2019. The paper explores the rationales that underpin resistance to formal grading in Norwegian education policy and practices in primary and secondary education.
The study addresses the following research questions:

1 What arguments were put forward for implementing and sustaining the policy of prohibiting formal grading in primary education and for reducing formal grading in secondary education?
2 What changes can be observed in the assessment policy discourses related to formal grading in primary and secondary education from 1939 to 2019?

The paper discusses how the transnational policy and research discourses on formative assessment and Assessment for Learning resonates with the Scandinavian countries' legacy of resisting formal grading.

Poster 23 Promoting job readiness for the 21st-century workplace. The empowerment of the Learning to Learn skill through the Assessment as Learning approach

Alessia Bevilacqua[1]
[1]*University of Verona, Italy*

In a constantly evolving labor market, the combination of technical skills and key competences can enable young people to adapt to changes maturing high levels of employability. Personal, social and Learning to Learn (LtoL) competence can be considered a relevant component of job readiness.
The empowerment of the LtoL skill at university is a complex process which involves meaningful changes. Assessment can become a useful compass to guide teachers in a renewal process. The Assessment as Learning (AaL) approach can be adopted to support students in the implementation of those meta-cognitive strategies which allow them to monitor constantly their learning processes and use feedbacks to make adjustments.
In the University of Verona, an experience of AaL has been implemented to support students coping with high-level cognitive processes. A convergent parallel mixed-method QUAN-QUAL research project was realized with two aims. The standardized questionnaire AMOS was proposed to verify the effectiveness of the AaL in promoting the empowerment of the LtoL skill. A descriptive focus group was implemented to understand the students' perceptions concerning the assessment strategies. Expected outcomes include benefits regarding students' learning outcomes, the strengthening of their transversal skills, as well as the individual well-being.

Poster 24 The effect of visual-to-verbal number transcoding on mathematics achievement

Dmitrii Kholiavin[1]*, Diana Kaiky*[1]*, Yulia Kuzmina*[1]*, Galina Larina*[1]
[1]*National Research University Higher School of Economics, Russia*

According to the Triple Code Model, numerical information can be manipulated and stored in three different formats: visual (Arabic digits), verbal (number words) and analogue representations. Numerical information can be transcoded from one format to another directly, which means that an individual may recruit verbal or analogue representations of numerosity to solve arithmetic problems presented in visual format. There is evidence that symbolic and non-symbolic magnitude representations and transcoding between them are associated with mathematics achievement. However, little is known about the relations between visual-to-verbal number transcoding and mathematics achievement.
The current study aims to estimate whether visual-to-verbal number transcoding associates with mathematics achievement taking into account symbolic number knowledge, phonological processing, reading skills and working memory. To fulfill our goal, we use data from 387 Russian first-graders (age varies from 7 to 8 years). Children completed tests on mathematics and reading achievement, working memory capacity, phonological processing, number recognition and visual-to-verbal transcoding. The expected results may be useful for theoretical and empirical support of the Triple Code Model of magnitude processing.

Poster 25 Building Inclusive Assessment Platforms

Luc Schomer[1]*, Sam Sipasseuth*[1]
[1]*Open Assessment Technologies, Luxembourg*

Technology is rapidly changing our day-to-day lives, including how we learn, teach and assess. Digitalisation offers all kinds of great opportunities, such as: innovative item types, assessment of 21st century skills like collaborative problem-solving, or improving efficiency of assessment. However, digitalisation also offers the opportunity to overcome barriers.
It is crucial to keep in mind, that not every test taker has the same individual capabilities to use digital assessment software. As defined by multiple directives (EU Web Accessibility

Directive, European Accessibility Act), a platform needs to be usable and accessible regardless of the users' age, educational background, or abilities. Therefore, digital assessment must be designed in a Human-Centric way, by being mindful of test takers' capabilities.
This poster will present our experience in creating our assessment platform, suitable for different age groups and accessibility needs.
We will introduce the open standards that form the basis for our work (W3C WCAG 2.1, WAI, IMS QTI 3) and show how Human-Centric Design enabled us to shape our open source platform together with our users.
We hope that our poster will inspire the audience to consider the test takers' individual capabilities and to tear down barriers by building inclusive assessments.

12.45 - 13.45 Lunch

Open Paper Sessions

Session A: Papers 1-3 – Psychometrics I
Chair: Amina Afif, Room: Castelo 1-2

13:45 - 14:15 On IRT models for analysing high tariff items
Yaw Bimpeh[1]
[1]*AQA, United Kingdom*

Many high-stakes examinations in the UK use both constructed-response items and selected-response items. The increased use of constructed-response items with high tariff response categories has motivated interest in polytomous IRT models. For example the General Certificate of Secondary Education (GCSE) English language examination paper consists of items that have a high tariff (e.g. 8, 12, 16, 24 marks), resulting in a large number of response categories per item. Furthermore, there is likelihood that category frequency of responses can be zero for some items. It is therefore inappropriate to apply an IRT model like the Partial Credit Model to such data. The purpose of this study is to examine the application of Samejima's Continuous Response Model (CRM) as a suitable measurement model for high tariff items.
We discuss the application of CRM to the high tariff items data, using 2018 GCSE English Language data. We compared the performance of CRM with the extended Nominal Response Model (eNRM). The empirical evaluation shows that the CRM has some advantages over the eNRM. The CRM method does not require calibration of large numbers of response category parameters per item.

14:15 - 14:45 Dimensionality in reading comprehension testing. An empirical validation of psychometric divisibility into reading processes
Michael Tengberg[1]
[1]*Karlstad University, Sweden*

The present study reports findings about construct validity in the Swedish national reading test in ninth grade. Based on a representative sample of 600 students, the study uses 1) confirmatory factor analysis (CFA) to investigate whether the theoretical construct underlying the test design can be empirically validated, and 2) principal component analysis (PCA) to identify prevalent latent variables in the test.
The Swedish national reading test in ninth grade tests four dimensions (reading processes) of reading comprehension. The division of items into reading processes plays a key role in the assessment of student results. Thus, high levels of construct validity and test reliability could be expected.
Findings from the study suggest that the internal consistency of reading process subscales is questionable. CFA, furthermore, reveals that the expected dimensions explains only a very

limited degree of variance in student results. The PCA displays that a single factor may be extracted, but that a large part of the items rather serve as noise with regard to dimensionality. Therefore, it can be questioned whether the Swedish reading test really tests the reading dimensions it intends to test, and whether student results can validly be assessed by dependence on their performance at different subscales.

14:45 - 15:15 Modeling certainty-based marking on multiple-choice items: psychometrics meets decision theory
Qian Wu[1], Rianne Janssen[1]
[1]KU Leuven, Belgium

Certainty-Based Marking (CBM) requires examinees to rate their degree of certainty when they select their single-best answer to a multiple-choice question. The obtained score on an item depends on the correctness of the answer and the indicated degree of certainty: the higher the indicated certainty, the higher the score for a correct response but also the higher the penalty for an incorrect response. The expected item score is maximized if examinees truthfully report their level of certainty. However, decision theory states that people do not always make the (rational) choice that leads to the optimal outcome due to varying risk preferences and/or the over- or underestimation of success probabilities. The present study therefore looks into the response behaviors of 334 first-year students of physiotherapy on six exams with CBM. Both response accuracy and the choice of confidence level (and corresponding points and penalties) are modeled by combining the Rasch model with prospect theory. It is shown that lower-ability students tended to overestimate their certainty levels, whereas higher-ability students tended to underestimate, even on tests with easy items. Female students showed higher accuracy rates and certainty ratings, but no significant gender difference were found regarding the mis-calibration of certainty levels.

Session B: Papers 4-6 – Educational Policy
Chair: Rebecca Hamer, Room: Castelo 9

13:45 - 14:15 Transforming assessment policy: towards a more socially just approach to policy design and enactment
María Teresa Flórez Petour[1]
[1]University of Chile, Chile

This paper aims at illustrating the way in which currently predominant means of constructing assessment policies are not responding to emerging horizons of social justice in education and, therefore, new ways of thinking policy development in the field are needed. The critical approach in connection to assessment policy is presented through the findings of a research project where the author used the National Curriculum Assessment System in Chile (SIMCE) as a case. These findings illustrate a vertical conception of policy design, where elites have the main voice while all other actors are portrayed as mere implementers of policy and as responsible for its success or failure. This conception is in tension with the democratic dimension of social justice. Additionally, actors from practice experience external assessment as a pressure device that decontextualises diversity, generating an injustice of recognition. To illustrate an alternative, the author refers to a second project, which involves a co-constructed large-scale assessment system in the Municipality of Valparaíso, Chile. The paper exemplifies how the principles and elements of this system are more in tune with a democratic and diverse approach to assessment policy development, generating better educational trajectories for students that counteract their current experiences of injustice.

14:15 - 14:45 Making the case for assessment in educational discourses
Mary Richardson[1]
[1]*UCL Institute of Education, United Kingdom*

Assessment preferences and policies in England are part of a framework where successful schooling is characterised by detailed forms of measurement, continual analysis of exam results and the reporting of grades. Within the public domain, I argue that a duality of assessment discourses (summative equals bad, formative equals good) has emerged and this will be presented as neither helpful nor true. The paper builds on existing theories of assessment to argue for a more nuanced view of what happens in schools with the aim of challenging the extent to which public discourses focused on assessment are true (or not). A brave attitude to educational transformation should help us to decide what we might expect from educational assessments and consequently, how best to enact changes which reframe our views of what is important in education and, we might reclaim the notion of assessment as something more than a test result.

14:45 - 15:15 Critical approaches in educational assessment
Graeme Findlay[1]
[1]*SQA, United Kingdom*

In 2018-19 a research project was undertaken at SQA to review the content of several practical courses at National 5, Higher and Advanced Higher levels in relation to the way these subjects are taught and assessed in England, Wales and Northern Ireland: the other three jurisdictions in Britain. The courses included in the research were:

- Fashion and Textile Technology
- Health and Food Technology
- Practical Cake Craft
- Practical Cookery

The purpose of the research was two-fold: the first to critically evaluate the current content as defined in course specifications and ensure it was fit for purpose, and secondly, to review the content against awarding bodies in Britain, to ensure that we were in line with current practice in terms of content delineation. The research activity was collaborative in nature and brought together SQA staff from both qualification development and from research and policy.
The purpose of our presentation is to focus on subjects that do not get as much exposure as the more academic core subjects, but play equal importance in terms of their contribution to the economy of the country, based on the skills developed and the vocations they support, and to share our findings.

Session C: Papers 7-9 – Test Development I
Chair: Lesley Wiseman, Room: Castelo 8

13:45 - 14:15 Spoilt for choice? Is it a good idea to let students choose which questions they answer in an exam?
Tom Bramley[1]*, Victoria Crisp*[1]
[1]*Cambridge Assessment, United Kingdom*

For many years, question choice has been used in some UK public examinations, with students free to choose which questions they answer from a selection (within certain parameters). In this paper we distinguish some different scenarios in which choice (or 'optionality') arises and explore the arguments for and against using optional questions. In particular we discuss i) whether having optional questions makes exams fairer or more valid; and ii) whether it is possible to discover if optional questions are of different difficulty and hence make statistical adjustments to students' scores that can allow for this. We conclude that unless there is a very good reason for allowing question choice it should be avoided.

14:15 - 14:45 White space in assessment materials – “space to think” or a “waste of space”?

Charlotte Stephenson[1], Bryan Maddox[2, 3]
[1]AQA, United Kingdom
UEA, United Kingdom
[3]Assessment MicroAnalytics, United Kingdom

In this paper, we explore the significance of white space in test paper design in the context of high stakes general qualifications in England. Exams should assess a students’ ability on a given construct (e.g., Chemistry). However, the influence of text layout on performance may threaten the construct validity of assessments. This research explored the effects of white space in question papers on cognitive processing – using eye-tracking methods – and respondent perceptions. Thirty-two students’ (aged 15-16) eye-movements were tracked as they completed two abridged AQA GCSE Chemistry papers: one with restricted spacing and one with enhanced white space. Eye-tracking data suggested that respondents took longer to complete questions with enhanced white space but also made more careful observations of the question content. Conversely, restricted white space was associated with shorter response times and more frequent re-reading of item content. Interview data revealed that students preferred papers with enhanced space, reporting that the additional space made them feel calmer and papers were easier to read. These findings suggest that the amount of white space in assessments impacts on measurable differences in assessment response processes and respondent preferences. We discuss the implications of these findings for increasing validity in assessment design.

14:45 - 15:15 Establishing effective test length and cut-scores for formative assessment using informative Bayesian hypotheses

Hendrik Straat[1], Anton Béguin[1]
[1]Cito, Netherlands

In individualized learning trajectories, administering small tests focusing on a specific learning outcome to determine mastery of the learning objective and to evaluate whether a student can progress to other learning objectives is valuable for improved teaching and learning. For this type of application, testing time competes with direct learning time, and a large number of learning objectives could invoke a potentially large burden due to testing. Mastery must then be assessed on a limited number of items. This results in a trade-off between the accuracy of the mastery decision and the applicability of this type of formative evaluation in practical situations.
In this presentation, we evaluate empirical test data based on fine-grained learning objectives to establish suitable test lengths and cut-scores assessment given item characteristics and specificity and sensitivity. Response patterns are evaluated using Bayes Factors comparing the posterior distributions in line with mastery and non-mastery.
Given varying item characteristics, we explore how to assess the probability of providing the correct answers for mastering and non-mastering students. Then, we investigate different sets of informative Bayesian hypotheses putting constraints on the item probabilities for mastering and non-mastering students.

Session D: Papers 10-11 – Higher Education
Chair: Elena Papanastasiou, Room: Castelo 6-7

13:45 - 14:15 Feedback of the external assessment of Higher Education students on the subject of Portuguese Language

Patricia Engrácia[1], João Oliveira Baptista[1]
[1]DGEEC, Portugal

In this paper, we present the main results of a statistic study carried out by DGEEC on the performance in the Portuguese Language exam of students that started their Higher Education in 2016/17.
On Portuguese Language, the external assessment is the national exam, done at the end of secondary education – the same for all students on the final secondary school year. This study analyses the grades on the Portuguese Language exam carried out in the academic year 2015/16, the year immediately prior to the entry of these students in Higher Education.
The objective of this study is to provide a diagnosis of the Portuguese Language preparation of students at the time of entering Higher Education in different courses, institutions and scientific areas. At the same time, the aim is to understand how the Portuguese Language preparation of these students varies according to their demographic characteristics, regional origin and previous secondary education course.
Since the national Portuguese Language exam is a mandatory summative external assessment at the end of secondary education, this data presents an opportunity for a reflection on this kind of assessment.

14:15 - 14:45 The transformation of University admissions practices in England and its impact on A-level - are standards being maintained?

Rachel Taylor[1], Nadir Zanini[1]
[1]Ofqual, United Kingsdam

Most 18-year-olds in England apply to University ahead of sitting their A-levels. Applications are based on grades predicted by schools, and Universities make an offer of a place to individual students. Such offers are generally 'conditional' on students achieving certain grades in their A-levels. In recent years, however, there has been a rise in 'unconditional' offers – offers that do not depend on grades and essentially guarantee students a University place.
This transformation in University admissions practices has raised concerns that students with unconditional offers will be less motivated to study for their A levels and might therefore under-perform. This could have implications for qualification standards, given the approach used to maintain A level standards in England.
To explore this, our analyses consider the possible implications of unconditional offers for students' A level performance. Our analyses use regression techniques to control for background variables that might be related to performance (prior attainment, gender, ethnicity etc.), and we analyse data across multiple years to consider any changes over time as the number of unconditional offers has risen. The discussion focuses on the implications for maintaining A-level standards.

Session E: Papers 12-14 – Fairness and Social Justice
Chair: Stuart Shaw, Room: Castelo 4-5

13:45 - 14:15 Equity in education within the European Union; A study based on PISA 2015 data

Remco Feskens[1, 2], Cor Sluijter[1]
[1]Cito, Netherlands
[2]Twente University, Netherlands

Equity in education is a hot topic for international debate. In this presentation, equity across the educational systems of European Union member states is addressed by using PISA 2015 data. This is done by looking at the extent to which education in different countries can be considered inclusive and fair. In this presentation, inclusive education is reflected in the proportion of students that are 15 years of age and are still in school as well as the proportion of students obtaining certain basic skills.
Fairness relates to how well countries manage to achieve education outcomes independent of the background characteristics of students. PISA 2015 scores in EU countries are compared with a set of background characteristics of students to sketch the state of affairs as fairness is concerned. This set comprises gender, immigration history, home language, age, and the socio-economic status of students. These are related through both univariate and multivariate models to students' PISA 2015 scores in science, mathematics and reading. Thus, light is shed on the extent to which countries minimise the effects of irrelevant background characteristics on learning outcomes; i.e. how fair their education systems are.

14:15 - 14:45 "Fair assessment" in a time of high-stakes testing and increasing student diversity in schools: The voice of three Chilean schools from a social justice perspective

Tamara Rozas[1, 2]
[1]University College London, United Kingdom
[2]Universidad de Chile, Chile

Assessment has been linked to social justice issues particularly those promoting equality of opportunities. However, international research reports negative effects of high stakes testing against traditionally marginalised groups, which raise the question about the contribution of this kind of assessment to social justice. Chile has a national test with high consequences, suggesting discriminatory effects for students. Moreover, recent educational policies encouraging equality and inclusion in schools were implemented, leading me to ask: To what extent it is possible to develop inclusive school projects in a context of high stake testing? What is meant by "fair assessment" in a context of increasing student diversity? The study employed a qualitative design based on a multiple case study of three Chilean primary schools with inclusive projects. Data collection methods included interviews, focus groups, and observations with different members of the school community. The results suggest that the notion of "fair assessment" reveals tensions with the current national assessment and provides opportunities to discuss the role of assessment in social justice in a context of high stake testing.

14:45 - 15:15 Assessment of students with special needs. Challenges, dilemmas and tensions between national regulations and teacher practices

Astrid Gillespie[1]
[1]Oslo Metropolitan University, Norway

Summative assessment of students with special needs are scarcely discussed in the literature regarding assessment. This paper concerns teachers' experiences with assessment in relation to students with special needs and the challenges and dilemmas they face in

their practice. The Norwegian regulations state that grades in school subjects and on external exams must reflect the goals and learning outcomes as stated in The National Curriculum. Students with special needs are not necessarily given access to these goals as they attend special needs classes in certain subjects. 40 lower secondary school teachers were interviewed regarding their assessment practices related to students with special needs. The findings suggest that they face dilemmas and challenges in their effort to motivate and enhance students' learning outcomes and at the same time make sure they are working in accordance with the school regulations. Ethical dilemmas in terms of exposing students to test where the content is mostly unfamiliar to them, due to their attendance in special needs classes are discussed. Tensions that occur when the school authorities' regulations don't cater for a differentiated assessment practice is also illuminated.

Session AA: Papers 15-17 – Assessment of Practical Skills
Chair: Stéphanie Berger, Room: Castelo 10

13:45 - 14:15 Evaluating the impact of qualification reform on students' practical skills
Stuart Cadwallader[1]
[1]*Ofqual, United Kingdom*

In England, science qualifications for 18-year-old students (A levels) have recently been reformed. Practical skills are now assessed in two ways: via questions in written examinations and by teacher assessment. The teacher assessed component is competency-based and graded as 'Pass' or 'Unclassified'. This is a separate grade that does not contribute to the student's primary A level grade (of A*-E). These assessment arrangements are intended to facilitate frequent teaching and learning of practical skills by allowing teachers greater freedom to integrate practical work within their lessons. However, some stakeholders expressed concern that the approach may have the unintended consequence of causing schools to deprioritise practical work and to instead focus on examinations.
Ofqual, the qualifications and examinations regulator for England, has therefore undertaken research to explore the impact of the reform on the teaching and learning of practical skills. This presentation will discuss a cross-sectional quasi-experiment involving 2,733 students from 15 university science departments. The 'hands-on' practical skills of students with pre-reform A level science qualifications were compared to those of students with post-reform qualifications. The findings suggest that, since the reform, students' practical skills have not declined in chemistry or physics, and may have somewhat improved in biology.

14:15 - 14:45 Re-heated meals: Revisiting the teaching, learning and assessment of practical cookery in schools
Gill Elliott1, Jo Ireland[1]
[1]*Cambridge Assessment, United Kingdom*

The place of practical cookery within school subjects in England has, in recent years, been debated as part of concerns about the nation's health and obesity. Cookery has been a school subject for over a century, but has only ever held a minority place in the curriculum.
In 2017 we surveyed teachers of practical cookery in schools, in a repeat of a survey first carried out in 2007. We asked them about the ingredients used and the skills taught in practical cookery lessons at school and also about the issues they faced delivering practical cookery teaching and assessment through the school food curriculum.
We have found that the nature of the products being taught in schools has changed, with less emphasis on sugary baked items than previously, which is consistent with the development of healthy eating initiatives and awareness. However, many of the issues surrounding the teaching of cookery skills in schools identified in 2007, such as insufficient equipment, lesson time and parental support, remain unchanged. In this presentation we will discuss the implications of this research and the role of practical cookery teaching and assessment in schools in the future.

14:45 - 15:15 Which factors play a role in explaining results on cognitive and practical skills in technology by 14- to 15-year-old students?
Lien Willem[1], Jan Ardies[2], Jaan Harnisfeger[1], Rianne Janssen[1]
[1]KU Leuven, Belgium
[2]Artesis Plantijn Hogeschool Antwerpen, Belgium

To evaluate the quality of technological education in Flanders, a national assessment took place in 2017 in the first stage of secondary education. The interplay between performance of students on both a paper-and-pencil-test and four hands-on assignments and some background characteristics like gender, attitudes of students and their parents was investigated. Results show that the performance on the written test and the performance on the four hands-on assignments are strongly related with each other. There is no significant difference between boys and girls for the hands-on assignments. Boys do perform better on the written test. Students who are more motivated for technology at school and are more interested in technology perform better on some of the hands-on assignments and on the written test. Also the home environment of the students is related to their performance: students whose parents are involved in technology perform better on three practical tests and on the written test. Students whose parents have a positive attitude towards technology have better results on one hands-on assignment and on the written test.

Session AAA: Papers 18-20 – Reliability
Chair: Anabela Serrão, Room: Castelo 3

13:45 - 14:15 A new standard setting procedure for competency based performances
Bas Hemker[1]
[1]Cito, Netherlands

Communication and collaboration are two of the abilities related to 21st century skills. Both are involved in the measurement of the ability to have a discussion, that was measured in the context of a national assessment in the Netherlands. For this purpose, competency based performance tasks were developed together with an observational tool, while scoring models were constructed to quantify the performance of the candidate. In order to give relevant feedback, standards needed to be set on the score scales.
There was a large number of challenges that excluded common standard setting procedures, such as the item-centered Angoff-type approaches. Also, the Body of Work method could not be used as evaluation of the performances was very time consuming. In order to meet these challenges, a new adaptive bookmark-related procedure was developed. As the procedure aimed for low rater inconsistency, it also helped gathering evidence to evaluate validity of the measurement instrument.
The new procedure builds a bridge between our current assessments to the demands of future assessments, as evaluation of performances and relevant feedback will become increasingly important in education. The whole procedure, from scaling, selecting examples and setting the standards will be explained in more detail in the presentation.

14:15 - 14:45 Maximising the reliability and role of expert judgement in standard maintaining to account for changes in student performance
Milja Curcin[1], Beth Black[1]
[1]Ofqual, United Kingdom

During standard maintaining in secondary school qualifications in England, examination work (scripts) on key cut scores is scrutinised by expert examiners to ensure that the new cut score performance standard reflects the same performance as the previous session. This is an important aspect of standard maintaining since the statistical methods (based on

replicating value added from session to session) cannot take into account changes in cohort performance not specified in the prediction model, e.g. changes in teaching quality. However, expert recommendations represent a small number of script judgements in a narrow mark range around the statistically recommended cut-score, and cannot meaningfully challenge the statistical recommendations.
Comparative judgement (CJ) may be a promising method for maximising the reliability of expert judgment about script quality. We report on the findings from pilots trialling rank ordering, paired CJ, and a hybrid rank order/paired CJ method, each with 6-10 expert judges and 10-30 judgements per script; and a crowdsourcing online paired CJ, with 40 judges and 12 judgements per script.
We evaluate the effectiveness of these methods, aspects of designs used, plausibility of outcomes, and the potential for inclusion in operational standard maintaining to ensure that genuine changes in performance can be recognised.

14:45 - 15:15 Applying different measurement theories to evaluate marker reliability in vocational assessments
Zeeshan Rahman[1]
[1]*City & Guilds, United Kingdom*

A variety of factors that exist in the assessment process, such as markers, questions and learners, can introduce unreliability or error in results given to learners. The larger the error, the less confidence we have in assessment outcomes. This ultimately compromises the validity of assessments and the reputation of assessment organisations. Research can help evaluate the reliability of an assessment, provide important information on its quality, and indicate how it can be improved. City & Guilds carried out several research studies to evaluate marker reliability, which primarily involved multi-marker studies where groups of markers were asked to mark the same learner scripts for vocational examinations such as functional maths & English, advanced mathematics, electronic communication, beauty therapy and make-up artistry. The marking was evaluated using different measurement theories i.e. classical test theory (e.g. mean score differences, score correlation, grade agreement), item response theory (e.g. many-facet Rasch measurement model) and generalisability theory (e.g. phi/phi lambda and analysis of variance). The aim was to compare findings based on different approaches with a view to exploring the benefits and limitations of these theories in investigating marker reliability. This paper aims to provide an overview of findings from this research.

15:15 - 15:45 Coffee break

Session F: Papers 21-23 – Formative Assessment
Chair: Karen Dunn, Room: Castelo 1-2

15:45 - 16:15 Practice of formative assessment in teacher education: case studies of Australia and Vietnam
Anh Duong[1]
[1]*The University of Sydney, Australia*

Students from the starting point have many ways to reach their targeted goals, but they need information along the journey on how they are going, where they is approaching and how to get there. These kind of information may come from assessment feedback from lecturers, peers or even themselves during their learning. The study presented here looks at the practice of formative assessment and understand how it was conducted at teacher education in Australia and Vietnam. The study employed mixed-methods including questionnaire, interviews, observations and focus groups. The results show that formative assessment was conducted effectively and efficiently with student-centered teaching approach, constructive feedback and engaging teaching pedagogies. Lecturers and students have had interactive activities with the

aim of shifting the learning forward. Lecturers in both countries have shared the learning intentions and success criteria, collecting learning evidence (by observing and questioning, etc.), eliciting the evidence to provide feedback to students. While Vietnamese lecturers and students believed self-assessment and peer-assessment could help improving performance, their counterparts in Australian have remained doubtable on these techniques' effectiveness.

16:15 - 16:45 Student perspectives on formative feedback as part of writing portfolios in higher education

Sarah Hoem Iversen[1], Zoltan Varga[1], Monika Bader[1], Tony Burner[2]
[1]Western Norway University of Applied Sciences, Norway
[2]University of South-Eastern Norway, Norway

Despite the crucial role that students play in formative assessment practices, student perspectives on such practices are relatively under-researched. Through a qualitative analysis of 128 reflection notes written by student teachers of English, we investigated the students' perceptions of formative feedback as part of portfolio assessment at two teacher education institutions in Norway. As such, this paper contributes to bridging the gap between research and practice. Students received peer and teacher feedback on assignments and wrote reflection notes during the semester. Findings show that students are positive towards teacher feedback and highlight the significance of teacher praise. Main objections raised against peer feedback concern the lack of constructive criticism. However, positive attitudes towards peer discussion groups suggest that they may be a more effective way of implementing peer assessment than formalised written peer commentary. Student reflections suggest that a failure to understand the task and the feedback is a possible hindrance to successfully revising assignments. Overall, students' positive attitudes towards the portfolio process, which includes multiple drafting, suggest that students in higher education would benefit from more opportunities to revise and resubmit their work, yet they need adequate practice in providing peer feedback, and interpreting and implementing feedback in general.

16:45 - 17:15 How summative assessments can be formative: using reading comprehension item data to inform teaching

Jemma Coulton[1], Anne Kispal[1]
[1]National Foundation for Educational Research, United Kingdom

High stakes summative assessments tend to be viewed negatively by the teaching profession in comparison with formative or diagnostic assessment. Research suggests that there could be widespread benefit from integrating formative and summative assessment, and that assessment data from summative assessments can be used formatively, if it informs future teaching and learning.
There are several barriers to using summative test data formatively, including lack of assessment literacy in the teaching profession (in particular the use of test outcomes to identify specific individual strengths and weaknesses); what test data is provided and in what format; and a general feeling of data overload in the teaching profession.
As part of the development of NFER's low-stakes summative assessments for 11-year-olds, we aimed to produce formative guidance for teachers, through conducting response coding. This guidance, based on the results of a large scale standardisation trial, includes information on patterns of performance, pupils' strengths and common misconceptions. The aim of this guidance is to overcome the barriers to using test data formatively outlined above.
Through the example of our reading assessments we outline the potential of summative assessments to provide information to teachers that can be used formatively to inform teaching and learning.

Session G: Papers 24-26 – Assessing Receptive Skills
Chair: Andrew Boyle, Room: Castelo 9

15:45 - 16:15 Reading and Test-Taking Strategies used in the TOEFL Junior Standard Reading Test: Evidence from Retrospective Think-Aloud Protocols
Dina Tsagari[1], Trisevgeni Liontou[2]
[1]Oslo Metropolitan University, Norway
[2]Department of English Language & Literature / National & Kapodistrian University of Athens, Greece

Recent research in the field of language assessment literacy - LAL (Inbar, 2008; Taylor, 2013) focuses on the learning classroom, and teacher beliefs about assessments and their uses among others. However, very limited research has investigated students understanding of assessment and student processes during their own performances on language tests.
The aim of the current research was to empirically investigate the effect specific reading strategies have on the nature and product of reading comprehension in the context of the TOEFL Junior Standard Reading Comprehension Test. The data was collected from 50 EFL students aged 11 and 12 years from a primary state school took part (in individual sessions) in retrospective verbal protocols reporting (in their native language) on the reading strategies they use in order to complete each set of reading questions included in the test.
The empirical data collected and analyzed for the purposes of this project as well as the lessons learnt during the process of cognitive strategy training undertaken in exploring young EFL learners' comprehension abilities stress that developing LAL requires understanding of both the testing processes of test-takers as well as fundamental principles of languages assessment.
The paper will round off with recommendations for future research.

16:15 - 16:45 Transformation during assessment: practice effects during the ESLC listening test
Andrés Christiansen[1], Rianne Janssen[1]
[1]KU Leuven, Belgium

In contrast with the assumptions made in standard measurement models used in large-scale assessments, students' performance may change during the test administration. This change can be modeled as a function of item position in case of a test booklet design with item-order manipulations. The present study used an explanatory item response theory (IRT) framework to analyze item position effects in the 2012 European Survey on Language Competences. In order to measure the item position effects, a logistic IRT model was constructed considering the interaction between the position of the item within the test and its difficulty. Item position effects were found for listening. More specifically, for a large subset of items, item difficulty decreased along with item position; this effect is known as a "practice effect" or "learning effect." Even though the effect size varied across items, languages, and countries, it was found along with all main tested languages (English, French, German, and Spanish) in the same level of proficiency.

16:45 - 17:15 Automated Scoring of Open-Ended Items for Reading Literacy Assessment in the Russian language
Maxim Skryabin[1], Alina Ivanova[1]
[1]National Research University Higher School of Economics, Russia

The automated scoring services can save costs associated with hand-scoring in large scale assessments. However, despite the fact, that researchers can use automated scoring in different subject areas, including English language arts, mathematics, and 21-Century Skills, there is still a lack of knowledge on what are the measures of automated scoring that work best for the reading literacy, especially on the languages, other than English.

To our knowledge, the current study is the first attempt of using automated scoring in reading assessment in the Russian language. The sample for this study included more than 1700 four-grade students. Their reading literacy was measured with 35 items, three of which were presented in the open-ended form.
To automatically score these items, the supervised learning methods for text classification were used. The classification algorithms were the logistic regression, random forests, and naive Bayes classifier. Using this approach, we found that we can accurately predict the binary score for three specific reading assessment items. The results of the current study show that it is possible to use the proposed methodology of automated scoring to establish reliable estimates for children reading literacy assessment in the Russian language.

Session H: Papers 27-29 – National Tests and Examinations I
Chair: Bas Hemker, Room: Castelo 8

15:45 - 16:15 Post 16 Qualification Reforms: Impacts on mathematics teaching, learning and assessment in England

Ben Redmond[1], Jennie Golding[2], Grace Grima[1]
[1]Pearson UK, United Kingdom
[2]Institute of Education, United Kingdom

This paper explores the impact that changes in policy context have on teaching, learning, and assessment of post 16, calculus-rich mathematics 'A-levels'. It reports evidence of how recent developments in England have shaped the experiences of students and teachers of reformed qualifications. It is important that post 16 mathematics education encourages deep engagement with mathematics in a way that is relevant to the kinds of 21st-century skills that students will need to progress into work or further study, including meaningful mathematical interaction with an increasingly data rich world. The reformed qualifications, introduced for first large-scale examination in summer 2019, aim to meet this objective. They are characterised by an aspirational curriculum, with emphasis on conceptual fluency as well as mathematical problem solving and reasoning. The reformed Mathematics A level also includes mandatory engagement with a large dataset using appropriate technology. We report on enactment, in particular around challenges that the reformed mathematics qualifications have placed on teachers' subject and pedagogical knowledge. We also explore how teachers are assessing the new curriculum with a focus on the uncertainty around the content and demand for both students and teachers of associated summative assessments.

16:15 - 16:45 Qualifications Reform in Wales: Opportunities and challenges for high-stakes assessment

Oliver Stacey[1]
[1]Qualifications Wales, United Kingdom

Education in Wales is undergoing a significant programme of reform with an innovative new curriculum being introduced in 2020. The new curriculum has provided Qualifications Wales (QW) with the chance to consider how 14-16 qualifications and their associated high-stakes assessments should be reformed in order to complement the new curriculum and more effectively serve the learners of Wales.
To inform the approach to qualifications reform QW conducted a series of interviews with leading experts in the field of educational assessment in the UK and Ireland and have also embarked upon a programme of engagement with key stakeholders in the system around 14-16 qualifications. This research explored the experts and stakeholders views on the main issues and concerns with the current examinations system and opportunities and challenges for high-stakes assessment associated with qualifications reform.
Some of the features identified were common with other high-stakes assessment systems,

for example managing the tension between innovation and risk within the assessment system. By contrast others were more peculiar to Wales such as divergence in qualifications and high-stakes assessments from a larger neighbouring country (England) and the potential issues this has for qualification portability.

16:45 - 17:15 The Yellow Wallpaper effect: The difficulties of moderating coursework
Stephen Holmes[1], Ellie Keys[1], Beth Black[1]
[1]Ofqual, United Kingdom

Coursework is thought to be a valid method to assess skills which are difficult to assess in written examinations. In England, coursework is marked by teachers and exam board moderators review a sample of work from each centre and decide whether the marks are accurate or an adjustment is needed. We carried out a study with different moderators looking at the same centre samples and reviewing different samples from the same centres for 8 components taken by 16- or 18-year-olds.
We found a range of moderator decisions for the same samples, and contrasting decisions on the samples from the same centre. Some disagreement likely stems from the difficulty of assessing this type of work, and some from differing moderators' beliefs about the purpose of moderation – re-marking, or confirming centre marks.
We saw some evidence that the range of topics/projects within a centre could limit candidates' opportunity to show their best work, and could even present difficulties for moderation, with some moderators judging the work alone and some not wishing to 'penalise' candidates for the centre's project selection. We reflect on the purpose and process of moderation, and the ability of coursework to promote the intended skills and educational outcomes.

Session I: Papers 30-32 – Comparative Judgement I
Chair: Mary Richardson, Room: Castelo 6-7

15:45 - 16:15 A framework for describing comparability between alternative assessments
Stuart Shaw[1], Victoria Crisp[1], Sarah Hughes[1]
[1]Cambridge Assessment, United Kingdom

The credibility of an Awarding Organisation is reliant upon the claims it makes about its assessments (including comparability claims) and on the evidence it can provide in order to support such claims. For example, for syllabuses with options, such as the choice to conduct coursework or take an alternative to coursework exam, there is a qualification claim that overall candidates' results are comparable regardless of the choice made. This presentation describes a study which sought to design a structure that can be used to evaluate comparability between alternative assessments. The study was undertaken in two phases. The first phase of the research focused on the development of a framework for evaluating comparability against a set of four standards as well as a separate recording form for capturing declared comparability intentions and how well these have been achieved. In the second phase of the study, the framework was piloted using two assessment contexts: on-screen and paper-based tests; and an Alternative to Practical and a Practical test. Outcomes from the pilot, using two experts engaged with the framework and form, are summarised in terms of the comprehensibility, usefulness and frequency of application of the comparability framework and recording form.

16:15 - 16:45 Why a Unified Approach to Language Scales Matters: The Case for Comparative Judgement

Rose Clesham[1], Sarah Hughes[1]
[1]Pearson UK, United Kingdom

In high stakes English language testing, global test-takers increasingly need their scores to have an agreed currency across a wide range of international jurisdictions and contexts. Universities, professional bodies and increasingly governmental migration departments need to be able to understand and trust the meaning of achieved language proficiency levels across a range of language testing agencies. A major issue is that different alignment methodologies can result in significantly different language framework equivalences. This research reports on the use of Comparative Judgement (CJ) as a method of establishing alignment between an established Language Framework, the PTE Academic's Global Scale of English (GSE) and the new emergent China's Standards of English Language Ability (CSE). This study has been a collaboration between both CSE experts in China and PTE GSE experts based across Europe. The results indicate that using CJ methodologies may provide a more equitable, reliable and manageable approach to international language alignment studies. This presentation will discuss the outcomes of the study and the implications for Language curriculum and assessment developers, universities, governments and policy makers.

16:45 - 17:15 Transforming the marking of extended responses through an understanding of complex judgment processes

Ayesha Ahmed[1]
[1]University of Cambridge, United Kingdom

Marking extended written responses in a way that preserves validity remains a major challenge for the assessment community. I will share outcomes of interviews with markers of extended responses. These resulted in the development of evidenced-based criteria in which different levels of performance are described in a way that is qualitatively different in terms of key aspects of the construct. The aim was to reduce uncertainty and the use of heuristics in marker decisions when applying the criteria, and to preserve validity by keeping the construct at the heart of the marking process. I will also discuss the approach of more and less experienced markers. Experienced markers report that they 'internalise' the marking criteria, but what does this mean for the validity of judgments? What happens when confidence in expert intuitive judgments overrides willingness to make a more deliberate analysis of the evidence in a response? Can we encourage markers to be more open to changing their minds when working through an extended response and will this result in more valid judgments? I will discuss how we might transform marking systems for extended response questions by using better mark schemes and through our understanding of how complex judgments are made.

Session J: Papers 33-35 – On-Screen Assessment
Chair: Lenka Fiřtová, Room: Castelo 4-5

15:45 - 16:15 Student engagement with on-screen assessments: A systematic literature review

Carla Pastorino[1]
[1]Cambridge Assessment, United Kingdom

Assessment delivery models that include at least one on-screen component have become increasingly common. One often-cited, learner-centred reason for adopting on-screen assessments (OSAs) relates to their potential of being more engaging for candidates who are assumed to participate in daily digital activities. However, it is possible that what is known about digital activities for entertainment may not apply for assessment tasks.

The question of whether OSAs are engaging for candidates and, critically, whether they are more engaging than their paper-based counterparts, remains open.
To answer this question, in this paper we describe the first stage of an ongoing research project, a systematic literature review. Its objective was to gain a better understanding of what is known about candidate engagement with OSAs and to provide the parameters for the design of subsequent behavioural experiments (phase 2). Initial results revealed that while the topic of engagement with on-screen materials has been abundantly explored with regards to learning, more and more nuanced investigations are required to conclude whether OSAs are more or less engaging. We also discuss the various ways in which engagement is defined in this context and the characteristics of OSAs that have been studied in relation to engagement.

16:15 - 16:45 On-screen assessments for young learners: Considerations for on-screen item type design and usage

Sanjay Mistry[1]
[1]*Cambridge Assessment International Education, United Kingdom*

Understanding how younger learners interact with on-screen assessments is essential to designing valid and fair on-screen assessments for this age group. A study was conducted to investigate how young learners (age 6-12) interact with different on-screen item types, in terms of cognitive and motor skills and to inform design and functionality considerations.
The methodology was based on two phases of observational research of on-screen tests, using past paper-based assessment content. Phase 1 consisted of remote observations conducted by teachers in India, Indonesia and UAE. Phase 2 involved a validation of the findings generated in phase 1 through face-to-face observations of learners conducted in India and Indonesia. Key findings indicated significant interactions between learner age, learner region and their ability to complete on-screen item types, raising several key issues pertaining to how socio-cultural differences in learners' exposure to technology and developmental differences in cognitive abilities impact the validity of using on-screen assessments across differing contexts and age groups. Further research will include item prototyping with young learners and exploring in detail on-screen item type design and functionality.

16:45 - 17:15 The use of touchscreen vs. standard devices for marking high-stakes exams

Sarah Hughes[1], *Martina Kuvalja*[1]
[1]*Cambridge Assessment, United Kingdom*

In addition to other upgraded and some new features, the new RM Assessor3 (RMA3) marking software enables the use of standard (non-touchscreen) as well as touchscreen devices for the purpose of marking exams. OCR is looking to use RMA3 for some of the selected exams in 2019 and the old version of the software will be completely replaced by RMA3 from June 2020.
Before moving to marking using RMA3, OCR wants to investigate whether marking on touchscreen devices affects the quality of marking and users' experience of marking (compared to marking on standard devices when using the same version of the RM software). The main aim of the study was to compare the quality of marking using RMA3 on touchscreen and RMA3 standard devices. An additional aim was to record examiners' experience of marking using these two modes of marking (touchscreen and standard). Findings will be presented in relation to the quality of marking at an exam and question level, and markers' experience of marking using standard and touchscreen devices will be discussed.

Session BB: Papers 36-38 – Test Development II
Chair: Marieke van Onna, Room: Castelo 10

15:45 - 16:15 Developing command terms for assessment of performance and creating in the performing arts
Rebecca Hamer[1], Christina Haaf[1]
[1]International Baccalaureate, Netherlands

Students' focus on success makes assessment a powerful tool to steer learning when the desired learning outcomes are clear to students, teachers and examiners. Expected standards of student achievement can be communicated using action verbs or command terms, especially when these are clearly defined. However, many commonly used action verbs are derived using variations of Bloom's taxonomy for cognitive learning and often are not appropriate for describing achievement in the performing arts expressed through performing (e.g. playing an instrument or dancing) or creating (e.g. composition or choreography). Furthermore, when used, these action verbs often appear at different levels of learning complicating the communication of standards to students and teachers.
This study presents ongoing work on developing command terms linked to specific recognisable student attainment levels for performing and creating in music. Four attainment level vignettes were developed in collaboration with music teachers and member checked against teacher and examiner experiences of student work at the target age level. For each attainment level six to eight verbs were proposed, and teacher and examiner feedback were collected to derive definitions appropriate to musical performance and composition for a smaller set of action verbs/precursor command terms to assess music performance and composition.

16:15 - 16:45 Tests as texts: investigating test questions from a sociolinguistic perspective
Filio Constantinou[1]
[1]Cambridge Assessment, University of Cambridge, United Kingdom

Assessment has the potential to transform teaching and learning. For this potential to be realised, the tools via which assessment is performed need to be understood in depth. One of the most commonly used assessment tools in education is the written test. To date, written tests have been investigated mainly as measurement tools or as socio-political constructs. However, they are neither merely measurement tools nor merely socio-political constructs. In the first instance, they are linguistic entities, or texts. In an attempt to illuminate this less recognised facet of tests, this study investigated written tests from a sociolinguistic perspective. The study was informed by sociolinguistic theory that suggests that the linguistic features of a text are not arbitrary but are dictated by the situational context of communication (e.g. who is writing, for whom, for what purpose). Drawing on this theory, this research sought to understand the linguistic design of written questions both at a structural and a functional level. Specifically, it aimed to (a) identify the most common linguistic features of written questions and, (b) explain their prevalence by reference to the situational context. This presentation will report the findings of the study and discuss their implications.

16:45 - 17:15 Quality criteria for assessment design
Paul Newton[1]
[1]*Ofqual, United Kingdom*

From an assessment design perspective, the idea of 'assessment for transformation' seems doubly challenging, because it foregrounds both validity and impact. Yet, validity and impact are deeply inter-related concepts, and the ideal way to facilitate positive washback effects is to maximise the validity of the assessment procedure. This is the principle that underpins Measurement-Driven Instruction.
My presentation is intended to help assessment designers to build validity into their assessment procedures, by helping to explain what it means to do so. To achieve this goal, I have reconfigured the concepts of construct-underrepresentation and construct-irrelevant variance – the 'two major threats' to validity – with the support of a simple signal processing metaphor. The reconfigured concepts – signal deficiency and signal contamination – can be further transformed into quality criteria. These criteria – signal purity and signal saturation – then become the 'two major guarantors' of validity. They explain what needs to be achieved for an educational assessment – any educational assessment – to work well.

Session BBB: Papers 39-41 – Supporting Students' Performance
Chair: Gerry Shiel, Room: Castelo 3

15:45 - 16:15 Issues in using low-stakes assessment tools to identify and support at-risk students
Guri A. Nortvedt[1]*, Andreas Pettersen*[1]*, Anubha Rohatgi*[1]*, Karianne Berg Bratting*[1]
[1]*University of Oslo, Norway*

Adapted teaching is a steering principle within the Norwegian educational system (Lovdata, 2018), and in 2014 the current national mapping tests (to help teachers identify students in grades 1-3 at risk of lagging behind in numeracy) was introduced.
This assessment is part of a national policy to ensure equal opportunities for students to acquire the basic skills needed in further education. The mapping tests are considered low-stakes to students, teachers, and schools. The same tests are used five years in a row and are known to the teachers. Analysis reveals that, after five years, the tests still identify the same number of students as in the first implementation. Furthermore, item characteristics have not changed much. This might mean that teachers struggle to interpret test outcomes and plan interventions. A follow-up interview study supports the hypothesis that it is challenging for teachers to analyse test outcomes. Teachers also revealed misconceptions about assessment and at-risk students which may be counterproductive. The presentation will highlight outcomes from the follow-up study and discuss issues related to developing and using mapping tests to collect data to plan teaching activities for at-risk students.

16:15 - 16:45 Life gets in the way: Resits for students unable to present for high-stakes exams
Damian Murchan[1]*, Martyn Ware*[2]*, Robert Quinn*[2]*, Fabienne van der Kleij*[3]
[1]*Trinity College Dublin, Ireland*
[2]*Scottish Qualifications Authority, United Kingdom*
[3]*Australian Catholic University, Australia*

High-stakes examinations in upper secondary education play an influential role in the lives of students, teachers and parents. Results offer evidence of student achievement that help policymakers evaluate standards of learning and provide access to higher education and employment for students. Given the stakes, public trust is essential. Meeting this challenge is reflected in efforts to ensure that exam procedures are robust. This study explores another challenge: how to balance maintaining fidelity to administration processes with the rights of examinees to access assessments in a timely manner. Through no fault of their own, some

students may not be available at the time of test administration and have to wait long periods before another opportunity arises.
The study investigates the extent to which offering examination resits based on ad misericordiam arguments such as illness or family bereavement differs from more usual cases where resits are sought to improve a grade. The systematic review audits practice in a selection of high-stakes examination environments internationally, evaluating agencies' response to arguments made by students on compassionate grounds.
Revealing some opaqueness in procedure and diversity of practice, the study provides useful evidence to assessment professionals and policy makers to inform policy and practice.

16:45 - 17:15 Recommending learning materials to resit exam candidates using collaborative filtering
Eva de Schipper[1, 2], Remco Feskens[1, 2], Jos Keuning[1], Bernard Veldkamp[2]
[1]Cito, Netherlands
[2]Twente University, Netherlands

The ability to give feedback or feedforward is traditionally linked to formative assessment. However, summative assessment can potentially also give us a wealth of information on which to base feedback or feedforward. This paper explores collaborative filtering as a method of extracting such information from summative data.
After taking their secondary school exams, students in the Netherlands get the opportunity to do one resit. Reasons for a student to take this opportunity are to (a) improve their overall grade or (b) to obtain a passing grade for a course for which that is an obligatory secondary school diploma condition. Individualized feedforward based on their first exam could be useful for students that intend to do a resit. Due to time constraints (merely two to four weeks between the grading of the initial exam and the resit), the feedforward would have to be automated. We explore collaborative filtering as a way of providing such automated and individualized feedforward in a short amount of time. The aim is to provide exam candidates with learning materials – in the form of exam questions to practice with – that are tailored to their individual learning needs.

18.30 - 20.30 Events for members holding accreditation and for doctoral students
Location: Lisbon Geographical Society

Friday, 15th November

Session K: Papers 42-44 – International Surveys I
Chair: Beth Black, Room: Castelo 1-2

9:00 - 9:30 Transforming marking practice: the case of TIMSS 2019
Grace Grima[1], Mary Richardson[2], Tina Isaacs[2]
[1]Pearson UK, United Kingdom
[2]UCL Institute of Education, United Kingdom

This paper reports on the practice of marking the test items for the International Maths and Science Studies (TIMSS) in 2019 in England. In this test cycle, England was one of the countries to deliver and mark the tests on screen. This shift from paper to screen has involved changes in the planning, preparation and notably, the implementation processes. Two changes to assessment practice are considered in this paper: firstly, the use of fully electronic delivery and marking of TIMSS test items and second, the process of marking– entitled coding.
We present findings relating to coding where the majority of coders assessing TIMSS 2019 in England were not teachers, but undergraduates (studying maths and/or science) who

were recruited to complete the coding. Their experiences of coding were documented using Think Aloud Protocol observations and followed up by interviews to document individuals' experiences. The findings helped us to explore the issues inherent in substantial changes to practice from technical and personal perspectives. Ensuring that all participants in testing regimes are cognisant of the potential impact of moving from page to screen is critical to building shared, and more likely successful, models of assessment for the future.

9:30 - 10:00 The Test-Taking Behaviour of Irish Students in PISA 2015 and 2018: student engagement, interest, and concentration in computer-based assessment

Caroline McKeown[1], Sylvia Denner[1]
[1]Educational Research Centre, Ireland

This paper examines test-taking behaviour and strategies of students in Ireland participating in PISA in 2015 and 2018. Since 2012, students in Ireland participating in PISA have been asked to fill in a test-taking behaviour questionnaire directly after the cognitive assessment.
The nationally developed test-taking behaviour questionnaire focussed on the student experience of taking the test, how interested they were, and what they usually did if they did not know an answer to a question. The findings underline Irish students' relatively low familiarity with computer-based assessment, with 54 percent of students never having taken a test on computer before in 2018, compared to 57 percent in 2015. The analyses highlight strong levels of self-reported interest, engagement and task perseverance during PISA tests. Irish students also reported differential interest and ease of responding to the three core domains, with greater interest and ease reported for reading literacy items. Interactive science items were perceived by Irish students to be less difficult than non-interactive items in 2018.
The relationship between different test-taking behaviours and student performance in PISA 2015 is also outlined. The potential for process data to further inform understanding of test-taking behaviour, complementing student self-reports, is discussed.

10:00 - 10:30 Profiles of student motivation variables in grade-four TIMSS mathematics

Michalis Michaelides[1], Gavin Brown[2, 3], Hanna Eklöf[3], Elena Papanastasiou[4], Militsa Ivanova[1], Anastasios Markitsis[1]
[1]University of Cyprus, Cyprus
[2]University of Auckland, New Zealand
[3]Umeå University, Sweden
[4]University of Nicosia, Cyprus

Items measuring motivation and affect are regularly administered in the TIMSS Background Questionnaires. These variables have been shown to be positively associated with achievement scores. However, in models combining multiple predictors, not all of them are equally good predictors. Confidence in mathematics for example is a stronger predictor than enjoyment or value for the subject. In this study, a person-centered analytic approach was adopted: confidence in, and enjoyment of mathematics were used as input variables in cluster analysis using TIMSS 2015 data from grade-four students in 12 countries. Results indicated that some clusters consisted of students who scored consistently high, moderate, or low on both motivation variables; however, there were clusters with inconsistent ratings, e.g. students high on enjoyment but moderate in confidence. Systematic patterns appeared across the datasets: cluster mean achievement was associated with motivation; confidence and enjoyment ratings were usually consistent within clusters, but when not, confidence was more closely aligned with mean achievement; clusters differed in terms of gender and socioeconomic status. Cross-country results were quite similar. Findings suggest a positive association between motivation and achievement at the cluster level, but differential importance of confidence over enjoyment perceptions in this relationship.

Session L: Papers 45-46 – Assessing Mathematics I
Chair: George MacBride, Room: Castelo 9

9:00 - 9:30 MathemaTIC's Mini-Summative Assessments: Transforming Assessments in Digital Learning Pathways for Mathematics

Frauke Kesting[1], Carole Frieseisen[2], Filipe Lima da Cunha[2], Jacob Pucar[3]
[1]MENJE Luxembourg, Luxembourg
[2]SCRIPT - MENJE, Luxembourg
[3]Vretta Inc., Canada

Over the past five years, the Luxembourg Ministry of Education, in partnership with national and international experts, has developed over 500 interactive assessment-for-learning items designed to make mathematics relatable, tangible, and engaging for students. This assessment-for-learning resource is called MathemaTIC.
MathemaTIC's personalized learning environment contains learning pathways that are separated into different modules, each of which covers a key topic area that students work towards mastering. Each of the learning pathways is comprised of three different phases with unique item types: Learn, Practice, and Apply. Although the learning pathways and subsequent summative assessments proved to effectively engage students in their math education and provide teachers with an in-depth view of their students' performance at the end of each module, MathemaTIC has now introduced an additional layer of assessment items which will enhance both student engagement and performance monitoring.
This new layer of assessment is comprised of mini-summative assessments that are spread throughout the modules.
Through this open paper presentation, we will be highlighting the addition of this new assessment format that has been introduced into the MathemaTIC platform. These new assessment items are designed to assess students incrementally and more frequently throughout each module, while still remaining engaging and enjoyable.

9:30 - 10:00 MathemaTIC – using digital assessment to inform teachers of how students construct meaning in problem-solving

Amina Afif[1], Franck Salles[2]
[1]SCRIPT Data Division - Ministry of National Education, Luxembourg
[2]DEPP - Office of Student Assessment, Ministry of National Education, France

As education systems globally integrate technology in the classroom, educators are beginning to appreciate the power of digital assessments to transform teaching and learning. This is can be observed in the case of formative assessment of student learning where the resulting information shows to both the student and teacher, what the student knows, understands and can do. Today, technology facilitates exactly that by adding to the traditional paper-based testing, the dimensions of personalized tests, real-time tracking of learning strategies and rapid feedback on progress. This presentation will use a mathematical item for Grade 9 students, created in the MathemaTIC learning environment to demonstrate how data generated from an online formative assessment may help teachers better understand how students attempt to solve a mathematical task.
The interactive item, which requires students to solve a system of equations in a given real-life context, will be examined from a didactical point of view. Predictive score models will be used to illustrate the paths taken by students to solve the task and the resulting patterns that teachers can observe in their mathematical strategies used. Such data reports on the learning experience are very meaningful and can be used to re(shape) classroom teaching and learning.

Session M: Papers 47-49 – Students' "Voice" in Assessment
Chair: Sarah Hughes, Room: Castelo 8

9:00 - 9:30 Students as stakeholders in the development of new assessment systems: A case study from Scotland

Martyn Ware[1], Shakeh Manassian[1]
[1]Scottish Qualifications Authority, United Kingdom

Focusing on the theme of students as stakeholders in the development of new assessment systems, this presentation will describe work undertaken by the Scottish Qualifications Authority (SQA) and Young Scot, Scotland's national youth information and citizenship charity, to gather young people's views on the future of assessment. SQA believed it was important to gather the views of young people to complement those gathered from other sources as part of its wider 'Assessment Futures' programme of work. Over a nine month period between September 2017 and May 2018 and using an established and structured co-design process a group of approximately 10 young people from a range of educational backgrounds developed a shared view of what they believed were important features of a future assessment system. Drawing on their own experience and on an understanding of some of the key themes influencing the future of assessment, the young people summarised their findings in a report to SQA. This presentation will describe the rationale for SQA's engagement with young people and the process they followed. It will provide a summary of the key points of their report to SQA and describe how SQA is responding to the report's recommendations.

9:30 - 10:00 SQA Mental Health and Wellbeing Awards: meeting societal need with flexible learning and assessment

Jen Morrison[1], Elaine McFadyen[1]
[1]SQA, United Kingdom

The theme of this year's conference, 'Assessment for transformation: teaching, learning and improving educational outcomes' highlights the need for assessment to be instrumental in affecting change. As the principle awarding body in Scotland, SQA is committed to ensuring that our qualifications and assessment provision inspires, enables and encourages learning focussed on sustainable long-term holistic, societal needs.
The improvement of mental health is a fundamental priority for multiple agencies across Scotland and Europe. SQA is playing a pivotal part in the education of our children and young people in this area, proactively developing new qualifications to transform the approach to learning about mental health and wellbeing.
Whilst Personal and Social Education is part of Scottish curriculum, there is no comprehensive programme focused on mental health and wellbeing being delivered within our schools.
This presentation will introduce delegates to the new Awards that SQA has developed to begin addressing this gap in provision, and the enormous potential for learners to understand and improve their mental health through education.
We will discuss unique delivery and assessment approaches aiming to ensure inclusivity and open access to learners from a multitude of learning environments, and share information about the exciting future for this development.

10:00 - 10:30 Tracking test motivation in low-stakes large-scale assessment: the case of the National Reference Test (NRT) in England

Ming Wei Lee[1]
[1]Ofqual, United Kingdom

In England, the National Reference Test (NRT) started in the same year as the first examination of the reformed GCSEs. It examines the GCSE curricula in English language and mathematics and is sat annually by a nationally representative sample of 16-year-olds. By assessing

successive cohorts on the same test questions, the NRT provides an indicator of the nation's 16-year-olds' attainment in the two subjects, which over time can be used as a monitor of any transforming effect of the GCSE reform on educational outcomes. Because NRT results are not reported for individual students, the test is relatively low-stakes for test participants. The low-stakes nature means that for the NRT to provide a longitudinal indicator of attainment, there needs to be longitudinal stability in the level of test motivation and/or the relationship between test motivation and performance in the samples of test participants of successive cohorts. To monitor NRT-specific test motivation and post-reform teaching and learning for GCSEs, the NRT student survey accompanies the NRT. Drawing on three years' data, this paper will report on the psychometric properties of the survey, the longitudinal stability of test motivation in the NRT and the emerging transforming effect of the GCSE reform.

Session N: Papers 50-52 – Statistical Approaches to Assessment
Chair: Nico Dieteren, Room: Castelo 6-7

9:00 - 9:30 Rurality and educational attainment in Northern Ireland: A multilevel analysis

Gemma Cherry[1]
[1]*Queen's University Belfast, United Kingdom*

Educational inequalities are a persisting problem and despite the vast amount of literature and research dedicated to this topic, the macro influence of location receives little attention. In the context of Northern Ireland, high-quality research on this topic is non-existent and little is known regarding the importance of location for understanding educational inequalities or if location interacts with other factors already known to influence educational outcomes. This research uses a quantitative, multilevel approach to provide for the first time in the Northern Ireland context, information regarding the influence of urban and rural locations on primary and post-primary pupils' educational attainment outcomes.
The findings highlight that rurality interacts with pupils' gender and socioeconomic status to influence the score which primary pupils achieve in English. Pupils who attend rural schools are found to have higher mean English scores compared to pupils who attend urban schools, indicating a rural advantage. However, the interaction reveals that not all rural pupils are equally advantaged. Boys from lower socioeconomic backgrounds who attend rural schools are identified as particularly at risk of lower English attainment. Rurality is also found to have a significant influence on post-primary pupils' educational attainment in GCSE maths and GCSE English.

9:30 - 10:00 Measuring and Correcting the 'Sawtooth Effect' in a First Award

Elena Mariani[1]
[1]*Pearson, United Kingdom*

The 'Sawtooth Effect' is a pattern of change in outcomes associated with reform whereby performance on high stakes assessments is adversely affected when that assessment undergoes reform, followed by improving performance over time as students and teachers gain familiarity with the new test - as defined by the UK Office of Qualifications and Examinations Regulation. Previous research suggested a rank ordering exercise of scripts from before and after the point at which test familiarity appears to have been reached as a valid approach to detect 'Sawtooth'. For a first sitting script perceived quality may be lower than it might have otherwise been if the 'Sawtooth Effect' was not present. Researchers at Pearson have developed a two-stage approach to quantify and correct for this effect for the first award of reformed International A-Level units. We define 'Sawtooth' in terms of the probability of winning a paired comparison. In the first stage rank order data of reformed

and unreformed scripts is analysed with a Bradley-Terry model. If there is evidence of 'Sawtooth', logit measures are converted into probabilities to recommend grade boundaries that are consistent with different levels of 'Sawtooth'. This presentation will illustrate this method and show examples of its application.

10:00 - 10:30 Exploratory Factor Analysis of the 2018 British Columbia Student Learning Survey
Todd Milford[1], Victor Glickman[1], John Anderson[1]
[1]University of Victoria, Canada

In British Columbia (BC), all students in the publically funded K-12 school system are assessed at the classroom, Provincial, and National/International levels to measure student learning and understanding. At the Provincial level, all students are assessed in reading, writing, and numeracy skills in Grades 4 and 7 (ages 9 and 12) and also assessed in Literacy and Numeracy at Grade 10 (age 15) and Literacy at Grade 12 (age 17). However, the Ministry of Education in BC also collects additional student information associated with student school experiences. The main information source in this area is the Student Learning Survey (SLS), an annual province-wide census of Grades 4, 7, 10 and 12 students, their parents, and staff in public schools. Despite this effort, limited psychometric information is currently available associated with this measure. In this presentation, we describe the questions that make up the SLS for students in Grades 4, 7, 10, and 12, detail some of the psychometric properties (e.g., item level descriptive statistics, loadings, internal consistency) of the measure, and make recommendations for how the SLS might best be used to inform students, parents, staff, and the general public on the school experience in BC.

Session O: Papers 53-55 – E-Assessment
Chair: Yaw Bimpeh, Room: Castelo 4-5

9:00 - 9:30 Modeling the construct in a computerized performance-based assessment of ICT literacy
Georgy Vasin[1], Svetlana Avdeeva[1]
[1]Higher School of Economics, Institute of Education, Russia

Information and communication technology (ICT) literacy is a complicated skill, and measuring it requires an innovative performance-based assessment. In this paper, we present some of our findings from 5 years of development on a computerized scenario-based measurement of ICT literacy. The scenarios are important to engage test takers and adequately measure their performance, but they also create local dependency. We explore the structure of ICT literacy, comparing it with the ACRL ICT literacy framework, with a focus on structural equation modeling (SEM) and introducing local dependency variables to improve model fit. Our results provide validity evidence for the ICL Test and point towards interesting applications for some of the construct-irrelevant variance collected by the assessment.

9:30 - 10:00 The INVALSI computer-based assessment: psychometric challenges and opportunities in test design and score reporting
Marta Desimoni[1], Donatella Papa[1], Cristina Lasorsa[1], Rosalba Ceravolo[1], Antonella Costanzo[1], Angela Verschoor[2]
[1]INVALSI, Italy
[2]Cito, Netherlands

The National Institute for the Evaluation of the Education System (INVALSI) every year assesses all students attending grade 2, 5, 8, 10 and 13 in Italy. From the school year 2017-2018, INVALSI national testing program has undergone two main changes. The first one is the transition from Paper and Pencil to Computer Based (CB) Assessment for grade 8, 10 and 13

(from the current year). The second change is about the score reporting: CB test results are not presented only as numerical scores, but also as described proficiency levels, and individual feedback is given by INVALSI to students attending grade 8 and 13.
These changes in the INVALSI program provide new challenges and opportunities for psychometrics. For instance, the inherent flexibility of CBA administration implies the needs for interchangeable multiple test forms in order to ensure test security. Furthermore, content and construct validity, as well as the efficiency of individual scores, need to be guaranteed, in order to describe proficiency levels and to provide reliable and valid individual feedback. In the present work, our application of Rasch item-banking to deal with these issues in the INVALSI CBA will be described, and data from the INVALSI-2018 will be reported.

10:00 - 10:30 Transforming national examinations from paper and pen to an online mode of delivery: how easy is it? Egypt 2019 - a case study
David McVeigh[1]
[1]*Pearson Qualification Services, United Kingdom*

In Egyptian education, a learner's entry into university is decided by their result in a single set of examinations. The Thanaweya Amma – taken at the end of Grade 12 (ages 16-17) – determines access to tertiary education, and thus drive a learner's progression and career path. The aggregate Grade 12 GPA determines whether a learner will progress to university, and if so, which course they meet the entry requirements for.
In this presentation we focus on the processes/challenges involved in transforming the assessment experience for a million candidates from paper/pen to wifi-enabled devices within a condensed time period in a country ranked 170th out of 200 for internet connectivity. We cover work undertaken to support the development and delivery of a technology and evidence-based management system to enhance service delivery at the classroom level and improve the student learning and assessment experience.
We will provide an overview of the Education Minister's vision to date and explain how we are working with stakeholders across the Egyptian education system to make the vision a reality. This includes the NCEEE, the organisation responsible for the development of the Thanaweya Amma, Ministry of Education representatives, colleagues from Sayegh publishing group among others.

Session CC: Papers 56-58 – Assessment of Hard to Measure Skills
Chair: Tom Bramley, Room: Castelo 10

9:00 - 9:30 Measuring critical thinking through innovative assessment: An investigation of the dimensionality
Irana Uglanova[1]
[1]*National Research University Higher School of Economics, Russia*

Critical thinking is commonly considered as one of the key competences in the 21st century. In this study we present a new conceptual framework for assessing critical thinking at the end of primary school. According to our conceptual framework, critical thinking consists of two components: Analysis and Conclusion, and it is necessary to achieve a certain level of Analysis as a prerequisite to achieve a certain level of Conclusion. In order to assess critical thinking, an innovative assessment instrument was developed. The instrument includes three scenario-based tasks which simulate learning and everyday situations. The innovative format of the instrument allows to take into account the type of relationship between Analysis and Conclusion. The sample consists of 500 Russian fourth grade students (9–11 year-olds). Analysis was conducted via a Bayesian network. We applied the Posterior predictive model checking framework with Standardized Generalized Dimensionality Discrepancy Measure (SGDDM) as a discrepancy measure. According to the results, the

theoretically expected inhibitory relationship between Analysis and Conclusion was confirmed. The implication of the findings and challenges in critical thinking assessment is discussed.

9:30 - 10:00 Assessment of problem-solving skills

Martina Kuvalja[1], Stuart Shaw[1], Sarah Mattey[1], Giota Petkaki[1]
[1]Cambridge Assessment, United Kingdom

It is crucial to provide an accessible description of the theoretical construct(s) which underlie assessments. This is especially important for exams that attempt to elicit complex, higher-order constructs that are "hard-to-measure" (Stecher & Hamilton, 2014). If these construct(s) are not well defined and understood, then it will be difficult to support the claims we wish make about the usefulness of the assessments, including claims that they do not suffer from factors such as construct under-representation and construct irrelevant variance. This work focused on one such skill, problem-solving, and aimed to identify problem-solving processes and behaviours described in the literature and to explore how these are usually assessed. Two specific problem solving contexts are investigated: domain-general (cross-curricular problem solving for which a specific curricular knowledge is not required) and domain-specific (specific to a certain domain/ subject and requires a certain level of subject knowledge).
Different assessment models for assessing problem-solving skills are presented and analysed through examples from PISA and Cambridge Assessment International Education assessments. Validity issues associated with each model are discussed and the recommendations for assessment design are made in order to improve the authenticity of assessment tasks and, therefore, to minimise threats to construct validity.

10:00 - 10:30 Is search for explicitly stated information really a lower-order cognitive skill in reading comprehension assessments?

Inna Antipkina[1], Ekatertina Aleksandrova[1], Alina Ivanova[2]
[1]Higher School of Economics, Russia
[2]National Research University Higher School of Economics, Russia

Despite Russian 4th graders usually demonstrated high results in PIRLS assessments, some researchers raised concerns that it is a 'colossus on clay feet' (Zuckerman, Kuznetsova, Baranova, 2018), because PISA reading results of Russian 15-year-olds are below average. The hypothesis of Zuckerman et al. (2018) is that students are extensively taught how to deal with 'higher order' reading questions (such as reasoning and giving opinions) and are not exposed enough to relatively easier tasks (such as looking for certain information in the text). In this study, we developed a reading assessment tool for 4th graders on the framework which includes PIRLS framework. We did find that empirical item difficulty did not match theoretically predicted item difficulty. Items requiring to find a piece of information in the text were often more difficult for children than items aimed at analytical and reasoning skills. However, we found that it was due to scoring rules. Children had to select the answer in the text clicking on relevant sentences, and only few of them chose the exactly needed pieces. More often, children overselected or underselected information. Thus, searching for particular piece directly in the text requires many analytical skills including the ability to ignore extra information.

Session CCC: Papers 59-61 – Perceptions of GCSE
Chair: Thierry Rocher, Room: Castelo 3

9:00 - 9:30 Teacher interpretations of GCSE specifications: transformational knowledge in the classroom - studying a novel

Jenny Smith[1, 2]
[1]*Independent researcher, United Kingdom*
[2]*University of Hertfordshire, United Kingdom*

Teachers' pedagogic discourse determines pupils' understanding of school-subject knowledge. My case-study research analyses the concept of powerful knowledge, based in a critical-realist epistemology and a social-realist theory of knowledge, in the secondary school classroom in England, following recent changes to specifications. Changes to GCSEs focussed on 'powerful knowledge', promoting an emphasis on knowledge based on academic disciplines. Powerful knowledge enables pupils to access the means to judge knowledge claims and thereby challenge and influence society – the transformational power of knowledge – within a social justice agenda. I discuss the learning trajectories of 15 Year 10 pupils over 12 weeks, as they studied a novel for their GCSE English literature course. The focus on critical analysis and evaluation in the subject specifications should encourage teachers to engage pupils in a richer discussion of the text, but the grade descriptors and mark schemes limited the teachers' interpretations of subject content. The focus on 'training' pupils to recognise and reproduce a single interpretation of the text means the potential for access to powerful knowledge is lost and therefore higher-grade outcomes become unobtainable. The concern about GCSE outcomes dominated the pedagogic discourse for teachers and pupils, resulting in a limited learning experience.

9:30 - 10:00 Teacher perceptions and experiences of the Non-Examination Assessment component of GCSEs in Wales: An exploration of fairness within the context of curriculum reform

Rachael Sperring[1], *Kerry Jones*[1]
[1]*Qualifications Wales, United Kingdom*

Non-Examination Assessment (NEA) i.e. assessments that are not taken by all candidates at a set time, on a set day, in exam conditions, is a component of most current GCSEs in Wales. Following an independent review of curriculum and assessment arrangements, Wales is undergoing a period of curriculum reform. As part of an ongoing programme of work to consider how qualifications can support this new curriculum, Qualifications Wales conducted a series of focus groups with teachers to explore whether the current approach to delivering and assessing NEA is appropriate. Teachers of ten GCSE subjects were asked about their experiences of NEA in six different locations across the country. Amongst the many themes that came out of these discussions was that the combination of NEA and examination within a qualification allows for a fairer chance for learners to maximise their educational outcomes. However, discussions relating to the varying levels of control applied to qualifications in relation to task setting, taking and marking were implied to impact upon perceived and actual fairness across centres and subjects. This presentation will explore the concept of fairness in relation to NEA as an assessment method in the context of curriculum reform and related qualification development.

10:00 - 10:30 Unpacking the difficulty of GCSE Modern Foreign Language questions by combining subject expert ratings and objective item features
Tim Stratton[1], Nadir Zanini[1]
[1]Ofqual, United Kingdom

Modern foreign language GCSEs (French, German, Spanish) were reformed for first assessment in England in 2018. We were interested in identifying if specific changes to item features as part of the reforms had impacted difficulty at item level. One key change was the introduction of questions written in the target language, which was a concern highlighted by teachers.
We constructed a beta regression model to predict item difficulty (facility) from item features. Higher and foundation tier exam papers for listening and reading skills for all three languages from 2017 and 2018 were used. Item level data was obtained from each of the three exam board providing the assessment in both years, giving a total of 24 exam papers per language. Item features included those coded from the exam materials and subject expect ratings of linguistic features. Our statistical models accounted for a significant amount of variance in facility scores (ranging from ~0.4 to ~0.6). Analysis indicated that target language had relatively little impact on the difficulty of the majority of items. The key predictors varied somewhat between assessments but question type and subject expert ratings of vocabulary demand were consistently among the best predictors.

10.30 - 11.00 Coffee break

Discussion group 1, Room: Castelo 1-2

11:00 - 12:00 Reforming national examination systems: Assessing new competences emphasizing interdisciplinary learning, student collaboration and creativity
Sverre Tveit[1], Christian Lundahl[2]
[1]University of Agder, Norway
[2]Örebro University, Sweden

This discussion group takes the ongoing reform of the national examination system in Norwegian secondary education as point of departure for discussing principle challenges facing nation states' examination and testing policies with respect to the validity and reliability of assessments and opportunities and threats related to digitalisation. By 2020, the national curriculum in Norway will be revised to reflect new developments including increased emphasis on interdisciplinary learning, student collaboration and creativity. In 2018 a national committee chaired by professor Sigrid Blömeke was nominated by the government to oversee the curriculum groups' proposals for each subject and to recommend overall changes to the examination system accompanying the reform. Based on review of research, the committee proposed multiple changes to the examination system.
The discussion centres around the following questions:

1 How can national examination systems be revised to be coherent with curricula reforms emphasising new competences such as interdisciplinary learning, student collaboration and creativity?
2 What are the main obstacles and threats related to implementation and administration of new types of examinations?
3 What are the main opportunities and threats related to digitalisation of national examination systems?

Participants are encouraged to contribute with policy and research experiences from other European contexts.

Discussion group 2, Room: Castelo 9

11:00 - 12:00 Transforming Assessment: How can digitalisation of high-stakes assessment enhance social inclusion through improved educational outcomes?

Irene Custodio[1], Kevin Mason[1], Grace Grima[1], Ellen Barrow[2]
[1]Pearson, United Kingdom
[2]Pearson Education, United Kingdom

Social inclusion is at the heart of the United Nations 2030 Agenda for Sustainable Development, with a commitment to ensuring a quality education for all. High-stakes assessment outcomes are an integral part of everyone's education, with significant impact on students' lives. Understanding the drivers of inequity in assessment outcomes remains a key question, and it is important to consider how high-stakes assessment affects social inclusion. In many ways, technology can be seen as a vehicle for addressing issues of inclusion and potential bias. However, issues of familiarity and access to technology must also be considered for economically disadvantaged learners.
The complexity of social inclusion means that it is not easy to separate and analyse specific social groups without, at the same time, considering the impact on others. The systems that underpin social disadvantage are complex, and care must be taken to harness technology for assessment to ensure a positive impact on learner inclusivity.
The discussion group is of interest to, and will benefit from, input from professionals working in high-stakes exam settings as well as practitioners and researchers working in the area of e-assessment, and also policy makers involved in national transformations of assessment from pen-and-paper to digital formats.

Discussion group 3, Room: Castelo 8

11:00 - 12:00 Assess@Learning - digital formative assessment in classrooms: A European Project

Jannette Elwood[1], Kay Livingston[2], Patricia Wastiau[3]
[1]Queen's University Belfast, United Kingdom
[2]University of Glasgow, United Kingdom
[3]European Schoolnet Partnership, Belgium

Assess@Learning (A@L) is funded under the Erasmus Key Action 3 Policy Experimentation call. The project will investigate the impact of using a systematic toolkit targeted at students, teachers, school leaders and system leaders on the system-wide use of Digital Formative Assessment (DFA). The partners are: European Schoolnet (EUN), Brussels; The Ministries of Education in Estonia, Finland, Greece, Spain and Portugal; IRVAPP, Italy; and the Universities of Glasgow and Queen's Belfast. The study seeks to provide empirically grounded responses to two questions:

- Does the Systemic Toolkit increase DFA adoption and improve its implementation?
- What is the impact of DFA on students' social attitudes and how does this vary according to their socio-cultural backgrounds?

The analysis of the data collected during field trials and through base line and follow-up questionnaires will be enhanced by qualitative evidence from Country and Student Dialogue Labs and the use of qualitative research approaches to capture and identify unanticipated outcomes of implementing DFA.
This discussion group will:

1 introduce the A@L project;
2 seek discussion about existing practices in schools across Europe of formative assessment generally and digital formative assessment specifically; and
3 discuss methodological choices around dialogue labs for educational assessment policy research.

12.00 - 1300 **General assembly**
Chair: Jannette Elwood, Room: Castelo 1-2

13.00 - 14.00 Lunch

Session P: Papers 62-64 – Language Issues in Assessment
Chair: Isabel Nisbet, Room: Castelo 1-2

14:00 - 14:30 Secondary school foreign language qualifications in England through the lens of the Common European Framework of Reference for Languages (CEFR): are assessment standards too high?
Milja Curcin[1], Beth Black[1]
[1]Ofqual, United Kingdom

There is a perception in England that modern foreign languages (MFL) secondary qualifications (GCSEs) are graded more severely and are thus more difficult compared to other subjects. This is often cited as a reason for declining MFL take-up. Simultaneously, the absence of clear grade performance descriptors may make it difficult for users of these qualifications to understand what students achieving different grades can do, and whether this is appropriate. This study interrogated the nature of performance and assessment standards in GCSE French, German and Spanish through the lens of the CEFR in order to provide a focus for discussion of grading standards in view of internationally recognised descriptors.
Grades 4, 7 and 9 on 2018 tests were linked to the CEFR scale by panels of experts through content mapping; rank ordering of GCSE performances and CEFR-benchmarked performances (for writing and speaking); and 'standard linking' using the 'Basket Method' to rate items on the tests in terms of the CEFR levels (for reading and listening). CEFR-related performance standards/descriptors and cut scores were extrapolated from this.
We discuss the implications for interpreting GCSE MFL grading standards based on these results in the context of GCSE MFL assessments, teaching practices, and accountability pressures.

14:30 - 15:00 Modern languages qualifications in Northern Ireland: student and teacher perceptions of difficulty, grading and decision-making
Leanne Henderson[1], Janice Carruthers[1], Ian Collen[1]
[1]Queen's University Belfast, United Kingdom

Learner uptake of languages at GCSE and A Level in England, Wales and Northern Ireland is in perpetual decline. These negative trends towards language learning beyond the compulsory phase (after age 14) are often attributed to low levels of learner interest and the 'English is enough' thesis. However, there is growing evidence that, in addition to the low value ascribed to languages by some groups of young people, perceptions of languages qualifications as less accessible than non-language qualifications have a negative effect on uptake at an individual level and provision at the school-level.
This mixed-methods study, conducted in multiple strands, engages with students and teachers in Northern Ireland to gather both quantitative and qualitative data which provide insights into the landscape of post-14 languages curriculum and assessment. The findings show how students experience choice and motivation in relation to studying languages, the factors they consider in exercising those choices and the potential for both internal and external factors to act as restrictions on their decision-making. This study provides original data relating to perceptions of languages qualifications as high-risk options with concerns reported by both students and teachers about content, difficulty and grading.

15:00 - 15:30 The CEFR as an assessment tool for learner linguistic and content competence: assisting learners in understanding the language proficiency needed for specific content goals in the CLIL classroom

Stuart Shaw[1]
[1]*Cambridge Assessment, United Kingdom*

The construction of an academic language proficiency scale whose model of reference is the Common European Framework of Reference for Languages (CEFR) has clear implications for Content and Integrated Learning (CLIL) pedagogy. CLIL introduces a cognitive dimension not explicitly treated in the CEFR – 'using language to learn'. However, a descriptor scale for academic language proficiency is complex and multidimensional, to the extent that a functional description of academic language use inevitably introduces a range of factors: cognitive stage, general language proficiency, the processes and skills involved in mastering the specific curricular objectives of each subject area, as well as the processes and skills involved in learning in general. Neither can it be assumed that these processes and skills are the same across countries or cultures.
An example of how an academic language scale may be employed in the CLIL classroom is in the application of learning outcomes. Both the content subject and the language used as the medium of instruction are similarly involved in defining the learning outcomes.
The clarity of content and academic learning outcomes can be enhanced with references to academic CEFR descriptors. By way of illustration, a history lesson plan focussing on mediation activities is described.

Session Q: Papers 65-67 – Psychometrics II
Chair: Tim Oates, Room: Castelo 9

14:00 - 14:30 Equating by pairwise comparisons

Marieke Van Onna[1], *Tecla Lampe*[1]
[1]*Cito, Netherlands*

We propose a new equating method, using expert pairwise comparisons of dichotomous items. It is suited for equating purposes in the case of two disjunct tests, that have been administered to non-comparable populations. Experts compare pairs of items from either test, and indicate which of the two is easier. Comparisons of pairs of same test items are derived from Rasch scale difficulty parameters. A fitted Bradley Terry model provides parameters for all items on one scale. This allows for multiple ways of IRT-equating.
A bootstrap procedure was used to estimate the 90% confidence interval of the cut-off score on the new test. The procedure was piloted on two disjunct tests of Dutch as a second language. An IRT-equating method had been used before, providing a way of evaluating the new method. The point estimate of the cut-off score on the new test, resulting from the proposed method with pairwise comparisons, was lower than the existing cut-off score. However, the 90% confidence interval included the existing cut-off score.

14:30 - 15:00 Balancing between psychometric validity and content validity: the case of differential item functioning for gender in a national assessment of French as a foreign language

Koen Aesaert[1], *Jo Denis*[1], *Karen Van Renterghem*[1]
[1]*KU Leuven, Belgium*

To produce accurate measures on the relationship between gender and students' performance, national assessment studies mostly take into account the potential presence of differential item functioning for gender. To improve the psychometric validity of the test, items that are flagged for DIF are often not retained for further test validation and subsequent analyses. However, removing DIF items may affect the content validity of the

test. This study illustrates the use of a technique to protect both content validity and psychometric validity of a test, by transforming DIF items into new variables that can be used to control for DIF in the subsequent correlational analyses. The data from a national assessment on primary school students' (n=2098) reading and listening comprehension in French as a foreign language are used. IRT based generalized linear mixed modeling was used to explore whether gender differences in students' reading and listening proficiency change when DIF for gender is taken into account. The results illustrate that this technique leads to more accurate measures without decreasing content validity of the test, i.e., the effect size of the relationship between gender and students' reading and listening proficiency decreases, when DIF for gender is taken into account.

15:00 - 15:30 How can we use Item Response Times in the Low-Stakes Testing? Ideas on Reliability, Cross-National Comparability, and Responses Classification
Denis Federiakin[1]
[1]*NRU Higher School of Economics, Russia*

The test score reliability is a sample-dependent characteristic closely related to the degree of sample differentiating. As a result of such dependence, if the test had failed to differentiate the sample much, the reliability may drop down. This problem has occurred in one of the tests used in the SUPER test project. Attempts to balance the test content across China and Russia made the test of professional competences for Electrical Engineering (EE) major students to be apparently very difficult for the sample. This problem partly related to the low motivation of students participating in the low-stakes situation: we discovered a large amount of atypically rapid responses among both Russian and Chinese EE students. To take into account differences in response process we used IRT modelling of item response time. We used the Hierarchical Generalized Linear Modelling framework proposed by J. van der Linden to analyse the relations between the ability and the response times. We also used IRTrees techniques to analyze aberrant responses that don't involve problem-solving behavior. The results suggest that not all ways to classify the responses into aberrant and non-aberrant are equally effective.

Session R: Papers 68-70 – School Improvement
Chair: Angela Verschoor, Room: Castelo 8

14:00 - 14:30 Developing a framework for school level data driven decision making to improve student achievements
Pāvels Pestovs[1]*, Dace Namsone*[1]*, Ilze Saleniece*[1]
[1]*University of Latvia, Latvia*

Latvia is undergoing a nation-wide curriculum reform in general education. The reform's goal is to focus on the development of transversal skills, complex learning outcomes and deeper learning. In order to achieve reform aims and ensure successful implementation of the curriculum changes, it is critical to build school capacity for data usage in decision making related to teaching and learning. Decade of the research has called for better use of assessment data for learning and teaching, but nevertheless the actual implementation and use of complex set of data is rather poor and insufficient. This research aims to test the newly developed framework that is designed to enable schools to get from the stage of "having data" to "using data as a meaningful source of information to better teaching and learning". The framework covers such aspects as large scale student assessment data, student surveys, teacher performance in the classroom, and elements of school leadership practices. The research arrives at the conclusion that the developed framework increases the likelihood of schools being able to use data in a purposeful and effective way and design an action plan for student achievement improvement.

14:30 - 15:00 Performance evaluation in Nazarbayev Intellectual Schools: evidence from school inspections

Raigul Kakabayeva[1], Gulmira Zhailauova[1], Gulnar Kurmanbayeva[1], Olga Mozhayeva[1]
[1]Autonomous educational organization Nazarbayev Intellectual Schools, Kazakhstan

This study focuses on school inspection practices in five Nazarbayev Intellectual schools. In the period from January 2019 to April 2019, a scheduled inspection of Nazarbayev Intellectual schools was conducted. An inspection framework was devised to underpin this study and obtain data. Data for this study was collected using criteria and procedures within the framework to assess a broad range of directions such as the planning of the educational activities, governance and management, teaching and learning, staff capacity, access to education, school culture and collaboration with the community, the safety and welfare of the school and resources. The main approaches of collecting information during the inspection were lesson observations, document review in order to verify compliance with the requirements of normative legal acts of the Republic of Kazakhstan, legal acts of the Autonomous educational organization Nazarbayev Intellectual schools regulating educational and pedagogical activities of employees, and analysis of the students achievements. The findings of this work show that although the schools have more positive outcomes than areas of development, these areas affect school performance overall.
All the weaknesses established during the inspections were analyzed to make responsible decisions and action plans were created to help schools to improve.

15:00 - 15:30 Preparing for high-stakes assessment of aspirational curricula: the role of educative resources

Alistair Hooper[1], Jennie Golding[2], Grace Grima[3]
[1]Pearson, United Kingdom
[2]University College London Institute of Education, United Kingdom
[3]Pearson UK, United Kingdom

In this paper we explore the impact of educative mathematics resources for secondary school students on their preparedness for high stakes assessments and explore how they are benefitting from transformative aspirations of the reformed qualification. This study draws on longitudinal qualitative and quantitative data from 16 schools with varying characteristics (33 classes with their teachers and Heads of Mathematics), in addition to one school using the resources for the first time, where we used a case study.
Teachers were interviewed in each term, with Spring interviews focused on a targeted lesson observation of each class. Data were collected annually from student focus groups from each class, and all students in the study classes were surveyed each year. Baseline and end of study attainment data were analysed to assess progression.
Findings show generally above average progression from users of the resources and in some cases, outstanding attainment in their summative assessment, demonstrating the transformative potential of the resource use for supporting preparation for high stakes assessments. Moreover, the attitudes of the students to their mathematics experiences were unusually positive for 11-16 year-olds in England in comparison with recentevidence.

Session S: Papers 71-72a – International Surveys II
Chair: Anton Béguin, Room: Castelo 6-7

14:00 - 14:30 Learning for or learning from PISA? Developing a tailor-made training course for sustainable transformation of education and assessment towards 21st century functional literacy

Nico Dieteren[1]
[1]*Cito, Netherlands*

Young generations of learners have to be prepared for life skills in 21st century contexts. PISA has been an inspiration in many countries to reform their education to prepare their students for these life skills. This also requires a change in the way of assessment: from knowledge based to skills based; and not only for 15-year olds that participate in PISA.
In preparing students, teachers play an important role. We report on our experience in the development of a training course for teachers how they can prepare their students in functional literacy. The scope and main objectives of this training were broadened to improve teachers' knowledge, skills and practical competences in assessing functional literacy in context. Separate attention in the course has been paid to Reading literacy, Scientific literacy, Mathematical literacy and Global Competence.
We show the general structure and approach of such training, the main learning objectives, the materials and exercises used and evaluation of results of actual training sessions. It is not so difficult to let the real 21st century world context enter your assessment. Constructing context-rich items with high standards of validity, reliability and objective marking for high-stakes tests, like final exams: that is the real challenge!

14:30 - 15:00 Language effects in PIRLS 2016: Towards a more thorough analysis of differential item functioning

Yasmine El Masri[1]*, Joshua McGrane*[1]
[1]*University of Oxford, United Kingdom*

Differential item functioning (DIF) methods are often used in educational assessments, including international large scale assessments (ILSAs) to detect items that are potentially biased towards or against certain groups sharing a common trait such as gender, socioeconomic status, ethnicity, language.
Despite the adoption of rigorous translation and adaptation methods in ILSAs, many studies identified items that behaved inconsistently across countries and language groups. Most of the research examining DIF in ILSAs has been carried out on TIMSS and PISA. Our research examines the extent language versions (English, French and Arabic) of PIRLS assessments are comparable in terms of readability of passages and item difficulty and demands. We use PIRLS data to propose a more thorough and less biased approach for detecting potential inconsistency in item behaviour across groups in educational assessments. We first present the results of an in-depth analysis of language versions of PIRLS 2016 passages and items administered in England, France and Dubai. We then compare the results of this analysis with a DIF analysis carried out using Rasch modelling followed by Andrich and Hagquist's (2012, 2015) two-way ANOVA approach for detecting DIF. Discrepancies in the findings from the qualitative and statistical approach will be presented and discussed.

15:00 - 15:30 Transforming Teaching, Learning and Assessment though TIMSS

Elena Papanastasiou[1]*, Maria Evagorou*[1]
[1]*University of Nicosia, Cyprus*

Although not utilized enough for the transformation of teaching and learning, international large-scale assessments provide a plethora of data and information that could be utilized within and between countries. The purpose of this presentation is to describe the ways in which data from the Trends in International Mathematics and Science Study (TIMSS) could be utilized within an educational system for professional development purposes. This is the case of a series of teacher professional development workshops that took place in Cyprus, with the ultimate goal of improving the teaching, learning, and assessment in science education.

Session Y: Papers 73-75 – Validity and Validation
Chair: Caroline Jongkamp, Room: Castelo 4-5

14:00 - 14:30 Assembled Validity: The Case of ILSAs

Bryan Maddox[1, 2], Bruno D. Zumbo[3], Camilla Addey[4]
[1]Assessment MicroAnalytics, United Kingdom
[2]University of East Anglia, United Kingdom
[3]University of British Columbia, Canada
[4]GEPS, Universitat Autònoma de Barcelona, Spain

The Standards for Educational and Psychological Testing recognise that validity judgments are informed by multiple sources of theory and evidence about the characteristics of assessment data, response processes and consequences. The assessment literature often suggests that validity judgments take place in an ideal space of theory and argument (Kane, 2016). Drawing on Actor Network Theory (ANT) we present a revisionist perspective recognising the different uses and consequences of test scores, diverse actors and rationales for participation (Newton 2010; Addey and Sellar 2018). We ask, who makes validity judgments, with what evidence, on whose behalf, and for what purpose?
We argue that validation is 'assembled' by the network of actors and institutions who develop and implement assessment programmes. In our view, validation is not a single judgement, but a series of judgments made by different actors on the rationales, quality and consequences of assessment programmes. We illustrate our argument with examples of ILSAs from UNESCO's 'LAMP' in Mongolia and Laos, and the OECD's 'PISA for Development' in Ecuador, Paraguay and Senegal, with interview testimony and observational evidence. We reflect on what this tells us about theories of assessment validity, and the use of The Standards in international contexts.

14:30 - 15:00 Validation of the student selection system used for Nazarbayev Intellectual Schools

Aigul Jandarova[1], Zamira Rakhymbayeva[1], Aidana Shilibekova[1], Olga Mozhayeva[1]
[1]AEO Nazarbayev Intellectual Schools, Kazakhstan

Almost 10 years ago Nazarbayev Intellectual Schools (NIS) in collaboration with its strategic partners has developed a sophisticated assessment system, which starts from students' selection process, follows by monitoring of students' progress and criteria-based assessment, which is in turn, consists of formative assessment, internal and external summative assessments.
For effective management of educational process, ensuring the provision of high-level educational services, informing community about the condition of education, standardized and valid assessment results are necessary.
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' selected in 2013 and 2014 were collected. As quantitative data, students' Mathematics results during several assessment procedures from Grades 7 to 10 were used. Students' academic success in Mathematics was also analysed in the context of their ability level to master natural and math sciences, which were determined by Ability test results.
This study advances our understanding of a need for introducing changes into the existing student selection system considering the update of Kazakhstani secondary education curriculum, inner and outer demands from interested parties and planned external accreditation of the student selection system by competent foreign agencies.

15:00 - 15:30 Enhancing assessment validity through the use of animated videos: An experimental study comparing text-based and animated situational judgement tests

Anastasios Karakolidis[1], Michael O'Leary[2], Darina Scully[2]
[1]Centre for Assessment Research Policy and Practice in Education (CARPE), Ireland
[2]Dublin City University, Ireland

The heavy reading demands of some text-based tests, along with their restrictive nature in terms of the complexity of what can be presented as stimuli, can have serious implications for validity. This study examined the extent to which animations constitute a useful alternative to text for assessing complex skills. Participants in the study were randomly assigned to take either an animated or a text-based version of the same situational judgment test. The results showed that those who took the animated version of the test performed significantly better than those who took the text-based version. However, the effectiveness of animations in reducing construct-irrelevant variance that is attributed to language-related factors was less clear cut. Overall, animations were found to reduce the variance attributed to construct-irrelevant factors. The significance and implications of this research are discussed.

Session DD: Papers 76-78 – Comparative Judgement II
Chair: Saskia Wools, Room: Castelo 10

14:00 - 14:30 A Comparative Judgement Approach to the Large-Scale Assessment of Primary Writing in England

Christopher Wheadon[1], Daisy Christodoulou[1], Patrick Barmby[1]
[1]No More Marking Ltd., United Kingdom

Writing assessment is a key feature of most education systems, yet there are limitations with traditional methods of assessing writing involving rubrics. It is now well established that reliable and valid measures of writing can be derived from comparative judgement approaches. The approach presented here extends previous work on comparative judgement of writing by directly involving teachers in a large number of schools in the judging. To ensure quality control, the process incorporated a process of 'anchoring' that ensured that teachers could not artificially inflate their own pupils' scores. The approach was used to assess the writing of 146,135 primary pupils in England in 2018-2019. Overall, the results showed that a comparative judgement approach to writing incorporating anchoring shows promise in providing a fair and robust large-scale method to assess writing.

14:30 - 15:00 Moderation of non-exam assessments: a novel approach using comparative judgement

Lucy Chambers[1], Sylvia Vitello[1], Carmen Vidal Rodeiro[1]
[1]Cambridge Assessment, United Kingdom

In England, many high-stakes qualifications include non-exam assessments that are marked by the teachers rather than external examiners. Awarding bodies then apply a moderation process to bring the marking of these assessments to an agreed standard. Current practice requires moderation to be conducted at centre level, with one moderator per centre who builds up a holistic view of the centre's approach to marking. As each centre is only viewed by one moderator, this raises challenges with regard to holding the standard across centres - this is currently overcome using standardisation and monitoring procedures.
In recent years, technological advances have allowed electronic submissions of candidates' work (e.g., portfolios). This opens the door for novel ways of moderating that can move beyond the allocation of centres to individual moderators towards a scenario in which candidates' work is distributed across multiple moderators (without being bound by centre). Such new methods could ensure that the marking standard is consistently applied across centres.

This research investigated, using simulation, whether comparative judgement (a technique whereby a series of two pieces of work are compared side-by-side to generate a rank order of work) could offer a feasible, and potentially more efficient, alternative to the current moderation process.

15:00 - 15:30 Judges' considerations in assessing children's writing in a comparative judgement process
Patrick Barmby[1], Daisy Christodoulou[1], Christopher Wheadon[1]
[1]No More Marking Ltd., United Kingdom

The use of comparative judgement (CJ) to assess open-ended answers has attracted increased recent interest from researchers and policy makers. Facilitating this interest are technological advancements such as web-based CJ platforms. CJ is highly reliable and time-efficient for assessing skills such as creative writing. However, one criticism of CJ is that without explicit assessment criteria, how comparative judgements are made in terms of considerations made by judges is unclear.
This study explored judges' considerations when comparatively judging primary children's writing. Judges (primary teachers) judged in pairs and discussed how they decided which piece was better when presented with pairs of pieces of writing. A web-based platform was used to carry out the CJ process. Screen recording software was used to record the pieces of writing appearing for each judgement, and also judges' conversations using the computer's built-in microphone. 24 judges were involved in the study, each pair carrying out two judging tasks for different ages of children. The conversations were transcribed, and thematic analysis used to draw out the main emerging considerations. In this paper, we will present the five main themes emerging from this analysis, providing an insight into how judgements are made in a CJ assessment of writing.

Session DDD: Papers 79-81 – Educational Approaches to Assessment
Chair: Ayesha Ahmed, Room: Castelo 3

14:00 - 14:30 Adapting the Cognitive Abilities Test (CAT4) to support teaching and learning in Chinese classrooms
Bernadetta Brzyska[1]
[1]GL Education, United Kingdom

Individualised or personalised learning is still a relatively new concept for teachers in many Chinese schools. China has started moving away from passive and rote learning style to a more "active, problem-solving learning style to improve students' overall abilities to process information, acquire knowledge, solve problems and learn cooperatively" (OECD 2016). The theme of this year's AEA Europe conference focuses on assessment that is used for transformation – it is therefore timely to present our study showing the findings of the nationwide trial of a Chinese version of GL Assessment's Cognitive Abilities Test (4th edition) in China including a standardisation of 15,000 students, and how we aim to encourage the use of CAT4 for enhanced teaching and learning for improved educational outcomes in China. The paper will be of interest to educationalists and test developers alike, giving detail on the trial, the adaptation into Chinese, the barriers overcome through a new platform hosted in China.

14:30 - 15:00 What do student skills assessments tell us about performance gaps by gender?
Marianne Fabre[1], Lea Chabanon[1], Thomas Portelli-Tronville[1]
[1]Direction de l'évaluation, de la prospective et de la performance [DEPP], France

Evaluation of school systems and measurement of learnings have been considerably developed in recent decades in Europe. More and more national and international assessments are being implemented, in different disciplines, at different times of schooling. These evaluations do not all respond to the same need (diagnosis, results or comparison) and it is sometimes difficult to synthesize information from different sources over time.

Using student assessment data collected in France, we try to produce an inventory of the differences in performance between girls and boys throughout their school career, distinguishing disciplines.
For this, we have identified and standardized all the performance gaps between girls and boys found in each assessment conducted by the Ministry since the early 2000s and even before.
This exercise reveals that the differences in skills by sex are much more pronounced in reading than in mathematics and this throughout schooling.
The differences in mathematics, in experimental sciences for example, in favor of boys, tend to stabilize in middle school. On the other hand, the gaps are large in reading in favor of girls in primary school and do not tend to decrease in secondary school.

15:00 - 15:30 Two centuries of 'cram' – a history of cramming in UK educational assessment
Lydia May Townsend[1]
[1]Institute of Education, University College London, United Kingdom

The recent focus on examination in education discourses in the UK may leave the impression that concerns about assessment are a modern insight. This is a fallacy. Many of the concerns modern educationalists have about assessment can traced back hundreds of years. This work focusses on the last two hundred of those. It is a historical analysis using original documents that have not previously featured in academic study in assessment. Specifically, it focusses on the notion of 'cram'. 'Cram' is the process where students learn only what is needed to pass an examination, usually in the hours, days or weeks leading up to the examination. This 'cram' is purported to lead to: a narrowing of the curriculum; a focus on retaining knowledge only long enough to pass the examination; 'gentleman' who in middle age have no desire to learn; and, students being encouraged to develop morally reprehensible habits. It is an issue of significant political concern. By looking at historical understandings of 'cram', this work gives context to modern examination and assessment discourse.

15.30 - 16.00 Coffee break

16.00 - 18.10 Ignite Session and
16.00 - 17.00 Symposia

Ignite Session
Chair: Andrej Novik, Room: Castelo 1-2

16:00 - 16:10 Guidelines for formative tests in the classroom based on memory research
Desirée Joosten - ten Brinke[1, 2], Kim Dirkx[1], Gino Camp[1]
[1]Open University of the Netherlands, Netherlands
[2]Fontys University of Applied Sciences, Netherlands

A formative test is often used in education as a tool to prepare students for final tests as they have has shown to be beneficial for learning. However, there are no clear guidelines for designing such formative tests. Memory research on effective learning strategies in the past

decade, however, has yielded useful results that could inform teachers on, for example, the most optimal testing format and the optimal distribution between tests for learning. The link between research on formative testing and testing as effective learning strategy (i.e., retrieval practice or testing-effect) has thus far not been thoroughly made. In this study, research paradigms from memory research and formative assessment research are combined, which has resulted in concrete and well-founded guidelines for the design of formative tests to be used in the classroom.

16:10 - 16:20 How can we tell if it is valid? Using operational data to build an argument for validity

Kevin Mason[1]
[1]*Pearson, United Kingdom*

We report on the development of a validation framework for Pearson General Qualifications. The purpose of this framework is to allow us to communicate validity in a public-facing report, in a way that is both robust and accessible. The framework has been based on similar processes developed by Pearson colleagues in the US, which have been brought into the UK context.
We apply the framework to the sitting in June 2018 of reform GCSE (9-1) Mathematics as the most recent series for which a full set of data is available in order to understand how this reformed qualification has functioned in its first year. A range of evidence to support the validity claim is collated from business-as-usual processes. We identify strong support in this evidence for claims of validity, as well as areas where evidence for support is weak, and more could be produced in future, especially in ensuring there is no systematic bias against particular groups of candidates. This is work-in-progress, and feedback from colleagues in the assessment world be will be very much appreciated.

16:20 - 16:30 Reflex: A generic app for evaluation and monitoring of formative assessments

Hendrik Straat[1]*, Romy Noordhof*[1]
[1]*Cito, Netherlands*

Key to formative assessments is to collect information to improve learning and teaching quality. This information has the form of feed-up, feed-back and feed-forward. Teachers and students can be assisted by insightfully organizing data for evaluation of the current assessment and for monitoring learning progression results over time to improve educational outcomes.
In this ignite presentation, we introduce Reflex, a generic environment for collecting, evaluating and monitoring of formative assessments on a wide variety of skills. The first version of the app is collects data about reflection on students' collaboration within an educative escape room. In this environment, students respond to questions about their current collaboration skills. After completion, students reflect on their own performance and formulate actions as recommendation for their next collaboration. Teacher can view the student responses. Teachers and students can also monitor the student's learning progress on different assessments regarding the same skill.

16:30 - 16:40 Getting out of their heads – using concept maps to elicit teachers' assessment literacy

Martin Johnson[1]*, Victoria Coleman*[1]
[1]*Cambridge Assessment, United Kingdom*

Although it is a key component of teacher professional competency, there are concerns in the UK that teachers have only limited Assessment Literacy (AL). Teacher AL is a difficult concept to define, and evaluating it represents a challenge. Many evaluations have

considered it in a narrow sense, but it is more than simply the acquisition of assessment knowledge and related skills, since it implicates a teacher's beliefs and feelings about assessment that have been acquired over time.
We used a novel concept mapping approach to elicit AL with a group of teachers who were also examiners. We wanted to see how formal examining affected their learning about assessment and helped to transform their understandings of assessment, and how this influenced their teaching. In this presentation we will outline the method in broad terms and discuss how it gains insight into embedded professional knowledge in ways that other methods find difficult.

16:40 - 16:50 Accelerating Innovations in Technology-Based Assessment
Mark Molenaar[1]
[1]*Open Assessment Technologies, Luxembourg*

The world of education and assessment is changing rapidly. Technology offers many new opportunities, but keeping up with technological advancements and expectations from digital natives is a daunting task for institutions and vendors alike.
This ignite session will focus on how open source and open standards can be leveraged to keep up with these innovations and even help accelerate them. Open source allows for multiple parties to contribute, offering a separation of concerns and enabling a sharing economy.
Open standards like IMS Portable Custom Interaction (PCI) and Standard on CAT allow for seamless interoperability of content and components, Technology Enhanced Items (TEIs) and adaptive engines respectively. Many of these new standards have introduced more flexible approaches to allow for market innovations.
Get inspired by technological innovations and learn how your organization can be part of the Learning Environment of tomorrow. The future of education is Open.

16:50 - 17:00 Micro-Analysis in Large-Scale Assessment
Bryan Maddox[1, 2]
[1]*Assessment MicroAnalytics, United Kingdom*
[2]*University of East Anglia, United Kingdom*

This presentation describes the use of micro-analytic process data in large-scale assessment, the methods used, the evidence produced, and its benefits for assessment design and validation. Micro-analysis captures fine-grained observational data on assessment response processes, such as audio transcripts, video, eye tracking and emotion recognition (Oranje, Gorin, Jia and Kerr 2017; Maddox, 2017; Newton, 2017; Zumbo, 2017). The data is a transformatory source of information on assessment performance that is fast becoming the new normal in the design and validation of 'next generation' computer-based assessments. The presentation gives examples from real-life assessments, including: International large-scale assessments; collaborative problem solving; exam paper design; and formative instruction software.

17:00 - 17:10 Using PISA process data for evaluating the validity of self-reported test-taking effort and the impact of low effort on item performance
Hanna Eklöf[1], *Peter Fjällström*[1]
[1]*Umeå University, Sweden*

An issue in PISA is whether students give their best effort to the low-stakes PISA test. Post-test self-report is commonly used to assess effort, but it has been questioned whether this is a valid measure of effort, or rather a proxy for self-evaluation of test performance.
As PISA is now computer-based, novel opportunities for research on student test-taking behavior through computer-generated log data are available. In the current study, a GLMM was used to analyze the relationship between time on task (log data) in relation to a) overall test performance and b) self-reported test-taking effort. Results show that low performing students do not benefit from spending more time on more demanding test items while the opposite is

true for students reporting low test-taking effort. Findings are interpreted as preliminary support for the effort measure and that choosing to invest more time and effort on PISA items could improve test performance.

17:10 - 17:20 Enhancing Learning and Assessment Systems via Continuous Tracking of Practice Assessment Analytics & Personalized Resource Recommendations
Alina von Davier[1]
[1]*ACT, United States*

Learners preparing to take summative, high-stakes assessments such as The ACT College Readiness Assessment will typically use resources to review the knowledge and skills that are associated with the requisite academic subjects. Given the broad scope of these subject domains, learners would benefit by receiving targeted, personalized lists of recommended resources that align with their individually diagnosed area needs. In our work, we have created a Recommendations and Diagnostics (RAD) API that can be plugged into a learning and assessment system to continuously track a learner's practice assessment analytics and translate that into predictions of skill mastery. Using these predictions, we drive a recommendation engine that prioritizes areas of need based on ACT's Holistic Framework and delivers sets of tagged open educational resources for learners to review. We discuss our hierarchical model that is based on LLTM and uses Elo ratings. Also, we discuss the role of industry standards such as IMS Global Caliper and the Competency & Academics Standards Exchange (CASE) as part of our initial integration into ACT's free test preparation solution.

17:20 - 17:30 Do classroom assessment scores affect future academic outcomes?
Jennifer Vinas-Forcade[1, 2]*, Cindy Mels*[3]*, Martin Valcke*[1]*, Ilse Derluyn*[1]
[1]*Ghent University, Belgium*
[2]*Instituto Nacional de Evaluación Educativa, Uruguay*
[3]*Universidad Católica del Uruguay, Uruguay*

In absence of external assessment, teacher assessment in the classroom does not only fulfill the formative but also the summative function of educational assessment. The latter includes assessment accrediting student performance in view of moving to the next grade level. Trust in teacher assessment is compromised by teacher subjectivity. Studies show teacher expectations affect and may bias student achievement scores.
Using a longitudinal database tracking a Uruguayan national student cohort, we conducted two-level logistic regression analysis to understand how behavior and achievement scores assigned by teachers, together with other features of students' primary school trajectories, individual, family and primary school characteristics are related with students' success in their first year of secondary school. We found both the achievement and behavior scores, as well as past grade repetition experiences, are significantly associated with success in secondary school, even after controlling for individual, family and school characteristics. We also observed gender bias in behavior scores. The findings stress that - when used for accountability purposes and given its longitudinal impact - teacher assessment should best be supported by professional development and grounded in national achievement standards and scoring criteria.

17:30 - 17:40 What happens when assessments are digitalised?
Anna Lind Pantzare[1]
[1]*Umeå University, Sweden*

In a time when digitalisation is high on the agenda, the digitalisation of assessments becomes an important issue to study. What should be assessed and how? Does the construct has to be revised? How does digital assessments affect what is taught and how it is taught?

In Sweden, the national tests, with the aim to support a fair and equal assessment and grading, are to be digitalised 2022. At upper secondary school, the national tests are end of course tests but they are not decisive. The teacher are to use them in the grading as one, albeit important, information about the students' knowledge. The ambition with the digitalisation is that the national tests should become more effective and that digitalisation will raise the equality. The efficiency is often mentioned together with automatic scoring and that score reporting will no longer be needed. Other efficiency arguments used are that digitalisation will reduce the handling of papers and there will be much easier to adapt the assessments for students with special needs. The equality arguments are often the same.
In this presentation, I will raise some challenges and possibilities concerning digitalisation of the national tests in mathematics for upper secondary school.

17:40 - 17:50 Combining proficiency measurement and mastery evaluation
Anton Béguin[1], Hendrik Straat[1]
[1]Cito, Netherlands

Providing relevant information about students' performance is the key purpose of assessment. Improving the value of this information for teachers and students could support them in designing suitable learning trajectories and potentially could improve the ways students are taught. To be able to optimize the available information about performance of a student the design of the assessment should be carefully considered. In the current study we evaluate what the requirements are in content and number of items if a test needs to report both on an overall proficiency level and also needs to report about mastery on a number of sub-domains.
The relevant concepts and techniques are introduced and applied on an empirical example. Using estimates from the empirical data a simulation study is carried out and based on the results preliminary guidelines are given for the minimum number of items per domain. This evaluation is done both for linear fixed form tests and for adaptive forms of testing.

17:50 - 18:00 The importance of establishing the validity of assessments in educational experiments
Andrew Boyle[1]
[1]AlphaPlus Consultancy, United Kingdom

In this presentation, I note how the desire for evidence-based policy leads to the use of experiments to evaluate the effectiveness of educational innovations. In such experiments, assessments are often used as outcome measures – to judge whether and how much students have improved under some new learning approach.
Researchers and organisations carrying out such experiments have much to think about (effect sizes, recruiting samples, ensuring appropriate administration of assessments, etc.). However, my contention is that researchers do not consider the validity of assessments used in educational experiments enough.

This is untenable; it is contrary to statements in internationally renowned testing standards, which put the onus on test users to ensure validity. Further, understanding the validity of a test in an educational experiment is tantamount to understanding the nature of learning (and / or of expertise) in the domain which is the object of the experiment.
Thus, this matter is not trivial, and evidence-based practice in education will be stronger if it addresses fully and convincingly issues of validity in the assessments it uses as outcome measures.
I will exemplify these arguments in the presentation, and seek the views of European colleagues who may have had similar experiences.

18:00 - 18:10 Redefining Student Success
Tanya Kolosova[1]
[1]YieldWise Inc, United States

Student Success Profile is a blueprint for identifying the best learners. Student Success Profile development process uses results of tests intended to assess knowledge.
The analysis of these assessments results in a Student Success Profile that describes the differentiating knowledge and skills of successful students. Unfortunately, knowledge assessments are often inappropriately analyzed and it leads to wrongly assessed attributes of Student Success Profiles, incorrect inferences about strengths, gaps, and opportunities, and misleading recommendations on how to close these gaps. We developed innovative mathematical approaches, algorithms and software solutions that not only help to overcome the problems with analysis of knowledge assessments but also help to build Student Success Profiles in a fully automated and scalable way. Our solutions provide accurate and reliable information about the abilities of a student, her/his strengths, opportunities and gaps, and eventually create an accurate quantitative estimation of Student Success Profiles attributes. Using our solutions, schools, colleges, and universities can create Student Success Profiles in the general domain of education and students retaining.

Symposia

Room: Castelo 9

16.00 - 17.00 The rare but persistent problem of errors in examination papers and other assessment instruments
Irenka Suto, Paul Newton, Joanna Williamson, Sylvia Vitello, Nicky Rushton

This symposium is about understanding why errors occasionally occur in examination papers and other assessment instruments, and why error detection can be slow despite the numerous checks included in most construction processes. We draw upon research on error reduction in complex sectors such as medicine, manufacturing, the nuclear industry, and aviation. In recent decades, greater understanding of how and why errors occur in these domains has been credited with significant improvements in safety and quality, saving countless lives.

All three papers assume that most assessment instrument construction processes form a complex system, since many of the numerous latent conditions that influence human performance are difficult to identify and measure. We share the theoretical position that system-level failure engenders human failure, which in turn gives rise to manifested errors such as those that appear in assessment instruments. The first paper provides an introduction to the problem of occasional errors, the second paper focuses on human influence, and the third paper focuses on system-level influence. Aspects of causation and of pre-emptive action to minimize errors are discussed throughout the symposium. Together we argue that the educational assessment community could benefit greatly by adopting principles of best practice developed in other industries.

Room: Castelo 8

16.00 - 17.00 Large Scale Digital Examns in Dutch Intermediate Vocational Education: Lessons Learned
Marcel Claessens, Peter Hakvoort, Maaike Beuving, Marieke van Onna, Rolf Vegar Olsen, Cor Sluijter

In 2010 the Dutch government decided to make certain that candidates graduating for different levels in Intermediate Vocational Education (IVE) would meet minimal standards

for Dutch, numeracy/arithmetics and English. National exams were to be introduced instead of having schools develop their own exams as was formerly the case. The Dutch National Board of Tests and Examinations, CvTE, and Cito started working closely together to develop a system optimally fit for purpose. The aim was having IVE graduates enter the labour market better prepared than before and improving their chances in Higher Vocational Education.
In this symposium, attendees will be given detailed information on how CvTE and Cito achieved these goals. We show how improving the national level on certain subjects can be achieved through compulsory large scale educational testing. We discuss several aspects of the approach used to develop and maintain the system through three presentations. One aimed at the political and educational context; one aimed at the construction process and one of the methodology and psychometrics involved. The discussion will focus on the question to what extent the approach taken can be applied in other countries.

Room: Castelo 6-7

16.00 - 17.00 Developing, Analysing and Using: The Experience of the Scottish National Standardised Assessments and their Focus on Supporting Teachers
Sarah Richardson, Helen Claydon, Bethany Davies, Sladana Krstic, Anaghaa Wagh

The purpose of educational assessment is to understand where a learner is in their learning development at the time of assessment (Masters, 2013). This enables teachers to make informed decisions and purposeful adjustments to teaching in order to support learners in their forward learning pathway. In the context of the classroom, using assessment to precisely determine where students are in their learning is challenging, particularly when students are at significantly different levels. Through the transformational potential of technology, there are new opportunities to develop assessments tailored to the individual needs of learners. Computer adaptive assessments provide such an opportunity (Martin & Lazendic, 2018). This symposium will focus on computer adaptive assessment and how it can be used to support teachers, using the experience of the Scottish National Standardised Assessment (SNSA). Computer adaptive assessment enables items, or clusters of items, to be presented to learners in a way that matches their current ability. Research shows adaptive assessments improve measurement precision, and have positive effects on student motivation and engagement (Martin & Lazendic, 2018). This is due to assessments being neither too easy, resulting in student boredom, nor too difficult, resulting in student frustration. The SNSA are computer based, adaptive assessments designed and delivered by the Australian Council for Educational Research (ACER), along with partner organisations -Twig and SCHOLAR, to support teaching and learning. SNSA forms part of the Scottish Government's National Improvement Framework for Scottish Education and are aligned with Scotland's Curriculum for Excellence (CfE). SNSA assesses domains Numeracy, Literacy and Writing in Primary 1, Primary 4, Primary 7 and Secondary 3 (4, 7, 11 and 14 year-olds).
Uniquely, SNSA features a three-phase adaptive assessment design, delivered online and on-demand, and is accessible for all learners, including those with additional support needs.
The on-demand design allows teachers to assign assessment as part of their regular teaching activities. Online delivery ensures consistent, automated scoring and provides stakeholders with instantaneous feedback on student learning at an individual, class, and school level.
While SNSA is a large scale assessment, it differs from large scale assessment conventions in its focus on supporting teachers to identify gaps in student learning at the time of the assessment, rather than on accountability. SNSA is low stakes. There is no pass-fail and results do not determine future outcomes for learners. Moreover, while the trends emerging from SNSA are reported nationally (Australian Council for Educational Research, 2018) they are not published in the form of league tables. This approach highlights their utility as an effective tool for teachers rather than for ranking schools.
Training is also a key feature of the programme. Training is provided to support teachers not only to administer the SNSA but also to develop understanding of the assessment design and how data from their reports is interpreted to inform their classroom practice. As such, individual

reports provide information on how a learner has performed on all knowledge and skills assessed. A class-level report provides a diagnosis of the class's performance as a whole on all constructs that are assessed and a school level report provides information on overall results of learners.
This symposium comprises three papers, each of which focuses on a distinct component of the SNSA. The first paper explores the demands of developing test items to meet the specific requirements of the SNSA, namely for an adaptive assessment that focuses on informing teacher practices. The second paper addresses the approaches used for psychometric analysis of the cluster-based assessment design. The third paper, reports on how teachers use data from the SNSA, drawing on case studies within two schools.

Room: Castelo 4-5

16.00 - 17.00 Progression: Everyone is a Learner

George MacBride, David Morrison-Love, Jannette Elwood, Louise Hayward, Ernest Spencer, Kara Makara, Janine Barnes, Elaine Sharpling, Alex Southern, David Stacey, Jane Waters

Context:
CAMAU is a three-year research project commissioned by the Welsh Government and enacted through a partnership between the University of Wales Trinity Saint David (UWTSD) and the University of Glasgow. This research is designed to support the development and implementation of a national curriculum based in all curricular domains on concepts of learning progression rather than on statements of standards. CAMAU activity is set in the context of Welsh Government educational policy, including its clear commitment to subsidiarity of decision-making and to co-construction of the curriculum through the involvement of practitioners at every stage of the process.

Research focus:
This research project seeks to explore:
- reconceptualisation of assessment of learning as forward looking, identifying what is essential for future learning, rather than as summative accounts of what has been learned;
- structuring the curriculum in terms of 'What Matters' and descriptions of learning essential for progression rather than as statements of standards (Harlen 2015, Heritage 2008);
- further development of the Integrity model (Hayward & Spencer 2010) identifying conditions required for effective sustainable change in education, specifically extending the involvement of practitioners and learners (Lundy et al. 2011).

Further, the research process offers insights into effective partnership between universities in different jurisdictions and into implications of working in a bilingual context.

CAMAU Processes:
Over three years, the joint university team has been working with teachers, learners and policy makers to:
- identify in each curricular area what matters in learning to ensure that young people in Wales are educated to be active and informed participants in society;
- develop shared conceptualisation and understandings of learning progression in all six curricular Areas of Learning and Experience (AoLE);
- develop progression frameworks and associated descriptions of learning for each AoLE.

Five sources of information were used to inform the development of progression frameworks and descriptions of learning for each 'What Matters' statement in each of the six AoLEs:
1 a review of research literature on understandings and models of progression
2 examples of how progression has been conceptualised and structured in policy in other education systems

3 policy makers' understandings of how a new curriculum can be developed effectively and sustainably through co-construction
4 practitioners' perceptions both of what matters in their pupils' learning for effective progression and of the processes of development through co-construction
5 learners' understandings of progression.

Within the commitment to co-construction and partnership, a key feature of the CAMAU research approach is the building of capacity across communities. To this end members of the CAMAU team both worked with teacher development groups and participated as members of policy groups within the curriculum governance structure. In seeking data on practitioners' and learners' understanding, the universities have developed and modelled toolkits to be used by teachers to develop their research awareness and skills as they seek the views of learners and of colleagues on progression in learning.

The symposium:
CAMAU had reported in a 2017 AEA-Europe Symposium on Points 1 and 2 above.
This symposium explores issues related to points 3 through 5.
The first paper provides insights into the views of policy makers; the second into those of practitioners, both those involved in development and those involved in engagement and trialling; the third into the views of learners through an innovative learner toolkit designed for more independent use by teachers. In addition to these findings, this symposium explores critically our research principles and processes, identifying and discussing issues related to sustainable informed policy development. All papers note implications of our approach for the development of curriculum and assessment policy and practice in other jurisdictions.

19.00 - 23.00 **Conference dinner**
Location: Montes Claros Restaurant

Saturday, 16th November

Session T: Papers 82-84 – Assessing Mathematics II
Chair: Grace Grima, Room: Castelo 1-2

9:00 - 9:30 Assessing mathematics competence in Design and Technology: policy intentions and practical outcomes
Cesare Aloisi[1], Gemma O'Brien[1], Sarah Carter[1], Stephen Wooding[1]
[1]AQA, United Kingdom

This research investigates the immediate consequences of an assessment policy transformation in England, the introduction of compulsory mathematics skills assessment in the Design and Technology (D&T) Product Design and Fashion and Textiles examinations aimed at 17-18 year olds. These new requirements represent a fundamental change to the assessment. However, they are in line with recent policy attempts to increase pupil exposure to numeracy.
Using a mixed-methods approach, we seek to position the D&T reform within national and international contexts and consider its intended and unintended effects on the validity of the examination, focusing in particular on its impact on the type of cohort attempting this qualification. Preliminary entry data suggest that there was a small shift in the size and composition of the cohort between 2018 and 2019. Upcoming attainment data will be able to provide a clearer picture on whether this had any impact on student outcomes. The authors consider this research as a case study of the intended and unintended impacts of embedding numeracy requirements into assessments.

9:30 - 10:00 The Nordic student experience: How do students in Finland, Norway and Sweden experience instructional quality in Language Arts and Mathematics?

Astrid Roe[1], Marte Blikstad-Balas[1], Michael Tengberg[2]
[1]University of Oslo, Norway
[2]Karlstad University, Sweden

The present study investigates how lower secondary students in Finland, Sweden and Norway evaluate their Language Arts and Mathematics teachers. Several studies have emphasized the benefits of using students' evaluation when studying when studying teaching quality. In the present study we have employed a thoroughly validated survey, the Ferguson Tripod Survey, Ferguson, 2010, which was developed for an American context, and investigated to what degree Finnish, Norwegian and Swedish students are able to discriminate between different aspects of their teachers' instruction – and to what degree their responses provide new information about instructional quality that could be used to develop instructional practices further. Our main focus is on items or constructs which were given relatively high or low score and items that showed statistically significant differences between countries. We found that the response patterns in the three countries had many similarities, although some items showed significant differences between countries and between subjects. The study contributes to the knowledge of how students experience the teaching of Language Arts and Mathematics in the three Nordic countries at the lower secondary level and which learning-promoting activities occur more frequently and less frequently than others, and than expected.

10:00 - 10:30 Investigation of Heterogeneity in Mathematics Abilities across Compulsory School through Vertical Scaling

Stéphanie Berger[1], Laura Helbling[1], Martin J. Tomasik[1, 2], Urs Moser[1]
[1]University of Zurich, Switzerland
[2]University of Witten/Herdecke, Germany

The assessment of students' strengths and weaknesses, and their progress over time is the starting point for targeting teaching and learning to the students' individual needs. To compare individual assessment outcomes across school grades, a vertical scale is required. Besides, a vertical scale can also serve as a tool for investigating the heterogeneity of students' abilities within and across grades. In our presentation, we outline the development of a vertical scale based on item response theory methods for measuring students' mathematics ability from third through ninth grade. The scale was established based on several hundreds of mathematics items by means of a common item design and a concurrent calibration approach. By means of the scale, we investigate the heterogeneity of students' mathematics ability within the seven target school grades, and contrast the within-grade heterogeneity with the development of students' ability across compulsory school. Preliminary results from our cross-sectional study showed comparable heterogeneity of students' mathematics ability within the seven grades studied. However, we found a considerable overlap of abilities between school grades, even after controlling for gender and language spoken at home. Based on these findings, we discuss the implications of this heterogeneity for individualized teaching and individualized formative assessment.

Session U: Papers 85-87 – Psychometrics III
Chair: Guri A. Nortvedt, Room: Castelo 9

9:00 - 9:30 Looking beyond the test scores: Latent motivational profiling of teenage English language learners from four country contexts
Karen Dunn[1]
[1]British Council, United Kingdom

Test outcomes are only the tip of the iceberg when it comes to understanding the engagement of a group, or groups, of learners with their subject, and the opportunities they are accorded to realise their potential. This paper explores motivational profiles of students at varying levels of L2 language aptitude from four countries: Bangladesh (n=1518), Colombia (n=1479), Spain (n=1773), and Sri Lanka (n=1439). Data were collected as part of a wider project supported by ministries of education. All participants completed a multi-skill English language test, plus an eight-scale motivation questionnaire. Analysis was carried out using Latent Variable Mixture Modelling (LVMM) in Mplus8. Rather than assuming homogeneity across any of the observed groupings, this person-centred analytic approach classifies participants according to shared attitudes and performances derived from the data. This analysis provided a nuanced insight into similarities and discrepancies in motivational profiles for students who may, at face value, not have been differentiated by educators or policy makers. This study also demonstrated that the same distinctions are not applicable across all educational contexts and cultures. The proposed discussion will focus on the value of such insights in highlighting areas where educational and policy interventions could be of real benefit to learners.

9:30 - 10:00 Screening System for Professional Training Programs for Israeli School Principals: Development, Operation and Validation
Avital Moshinsky[1], David Ziegler[1], Lisa Levy[2], Revital Nachum[2], Itay Soudry[2], Anat Shirazi[3], Helena Kimron[2], Hani Shilton[4]
[1]NITE, Israel
[2]Avney Rosha Institute, Israel
[3]Ministry of Education, Israel
[4]The Open University, Israel

A system that screen applicants to programs that train school principals is applied in Israel since 2014.
This system was developed in response to inadequacies with its predecessor. Disadvantages of the original system included low variance in candidates' final scores and a sense that the assessments were not reliable. The main drawback, however, was the fact that not enough program graduates applied for positions as school principals after completing the training program.
The system relies on three different tools to gauge non-cognitive variables: a structured evaluation questionnaire filled out by each candidate's supervisor, a personal-biographical questionnaire, and an assessment center comprised of five behavioral stations. The use of assessment centers to screen job applicants – especially those competing for management positions – is becoming more common.
In addition to describing this relatively new screening system, the lecture will summarize the findings from an initial validity study aimed at investigating whether the screening system actually meets its goals.

10:00 - 10:30 Validity and Validation of Formative Assessment
Saskia Wools[1]
[1]Cito, Netherlands

For all assessments, validity is an important concern. The concept of validity has been developed mainly in the context of summative high-stakes testing. However, formative assessment becomes increasingly more popular. This paper discusses the concept of validity for formative assessment. When the argument-based approach to validation is used for formative assessment, the procedure would be similar as for the validation of summative assessment. However, the interpretation and use argument for formative assessment consists of inferences regarding a score interpretation as well as inferences regarding a score use. Score-interpretation inferences cover claims about students' performance from the instrument, while score-use inferences involve decisions on this performance and possible consequences in the learning process. As for the validity argument, two perspectives on the validity of assessments are distinguished, a measurement perspective and a functional perspective. The measurement perspective focuses on the accuracy and precision of scores as measures of some construct, and the functional perspective focuses on how well the assessment serves its intended purposes. Both perspectives are important when evaluating validity evidence, for formative assessment, the functional perspective is of central concern, and the measurement perspective plays a supporting role.

Session V: Papers 88-90 – International Surveys III
Chair: Jean-Pierre Jeantheau, Room: Castelo 8

9:00 - 9:30 What do international large-scale assessments tell us about high achievement in mathematics and science, with specific reference to Ireland and some comparison countries?
Vasiliki Pitsia[1], Michael O'Leary[1], Gerry Shiel[2], Zita Lysaght[1]
[1]Dublin City University, Ireland
[2]Educational Research Centre, Dublin, Ireland

In contrast to the extensive research on low achievement, research on high achievement is scarce. As a consequence, evidence, also referring to Ireland, suggests that high achievers' needs are not being met by national education systems. In order to assist the Irish education system in this direction, this study undertakes an in-depth investigation of high achievement in mathematics and science in Ireland. Specifically, Irish data from the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) since 2000 are juxtaposed with those of countries that performed similarly to Ireland on average. This longitudinal and comparative investigation of assessment data indicated that while Ireland seems to be meeting the needs of low achievers in mathematics and science, the same is not true for high achievers. Given the range of information that international large-scale assessments provide, a natural progression of this work is to investigate the particularities of these performance patterns by looking at predictors of high achievement in mathematics and science through the use of these data. Such research would inform relevant policy and practice in Ireland and thus, give students at the upper edge of the performance distribution more opportunities to achieve their potential.

:30 - 10:00 Relationships between 15-year olds' access to technology, perceived competence, autonomy and attitudes to ICTs, and their performance on PISA 2015 science in Ireland

Sarah Mc Ateer[1], Lynsey O'Keeffe[1], Gerry Shiel[1], Caroline McKeown[1]
[1]Educational Research Centre, Dublin, Ireland

Technology is an ever-increasing aspect of everyday life and recent educational policies emphasise a move towards greater integration of technology in teaching, learning and assessment. This paper examines relationships between several aspects of ICT and students' science achievement in PISA 2015. Results for Ireland, OECD averages, and comparison countries were examined focusing on school type, gender, socio-economic status, and DEIS status. Students reported greater interest in ICT, perceived ICT autonomy and competence than across OECD countries. Perceived ICT autonomy and competency had a positive correlation with science performance. Students reported lower availability of ICT at school than on average across OECD countries, and were less likely to use ICT in school and at home for schoolwork. A hierarchical linear model indicated a negative relationship between students' ICT competency and science performance, which is in contrast to the initial results. In conclusion, the results highlight that further research is needed to investigate the potential ways in which ICT can be used to enrich teaching and learning, notably in science subjects, to consider the appropriate level and format of digital technology integration in classrooms, and to consider the framework and construct of ICT variables in future educational research.

10:00 - 10:30 The ability to read numbers: A universal measure?

Elena Kardanova[1], Dmitrii Kholiavin[1], Peter Tymms[2], Christine Merrell[2]
[1]National Research University Higher School of Economics, Russia
[2]Durham University, United Kingdom

It is proposed that there is a single pathway for the order in which children learn to identify numbers. Although a broad pathway is unsurprising, systematic variation might be expected because of teaching, language of instruction, age, country of origin or other factors. This study using data from the UK and Russia presents evidence that such variations are minor; when learning to identify numbers children largely follow the same pattern regardless of the country. This finding is important in two ways: (1) furthering our knowledge of children's early mathematics development, suggesting a scale which can inform teaching and learning, and providing a structure within which mathematical development can be anchored; (2) it provides a universal scale against which valid international comparisons of the mathematical development of young children can be made.

Session W: Papers 91-92 – Policy
Chair: Paul Newton, Room: Castelo 6-7

9:00 - 9:30 The 'grey history' of assessment: understanding the origins of England's new model of assessment of practical work in Science

Tim Oates[1]
[1]Cambridge Assessment, United Kingdom

England recently has introduced into its high stakes assessment a new model of assessment; one in which marks from practical work no longer contribute to the grades in the qualifications. This model caused considerable controversy, and was adopted by the national regulator in the midst of highly adverse reaction. Various organisations predicted a collapse of practical work in schools. However, initial piloting work suggested the opposite was occurring in the limited trial centres. On national roll-out, similar benefits seem to be occurring across the system. Critics continue to feel that practical work should contribute to grades, despite the positive findings of the initial and continuing evaluation studies.

The presentation will trace the history of the development of the new model, examine emerging evaluation research on its success and impact. In doing so it'll reveal aspect of 'grey' history not present in the official record. It will examine issues of context and background; why a radical and seemingly unpopular model was conceptualised and introduced. The analysis of the development and introduction of the new model gives insights into public policy-making and technical issue of assessment design. We believe key elements support international comparative work and national policy formation across many nations

9:30 - 10:00 Irish Primary Teachers' Use of and Attitudes to Standardised Achievement Testing in English Reading and Mathematics

Zita Lysaght[1, 2], Deirbhile Nic Craith[3], Michael O'Leary[1, 2]
[1]Dublin City University, Ireland
[2]Centre for Assessment Research Policy and Practice in Education (CARPE), Ireland
[3]Irish National Teachers' Organisation, Ireland

Recent policy changes in Ireland means that primary schools are required to administer standardised tests in English reading and mathematics in 2nd, 4th and 6th classes, and to report the aggregated results to their Boards of Management and the Department of Education and Skills (DES). Schools are also required to share the results with parents. As of September 2017, the results are used at national level as part of the process involved in determining the allocation of special educational teaching resources to schools. The international literature suggests that when standardised test scores are shared widely and used for purposes beyond internal planning, the associated sense of accountability can result in pressure to perform, narrowing of the curriculum and other negative consequences. This paper presents the key findings with respect to how a random sample of 1,564 teachers felt about standardised tests and how they were using them in schools. While the data indicated that teachers were neither wholly supportive, nor wholly opposed to standardised testing, there was also evidence that test data were underutilized for formative purposes and that the process of constructing standardised tests as well as the interpretation of norm scores were not well understood by many teachers.

Session X: Papers 93-95 – Assessment and Teachers´ Practice
Chair: Roger Murphy, Room: Castelo 4-5

9:00 - 9:30 Corpus-based teaching practices & classroom-based assessment: Putting theory into practice

Trisevgeni Liontou[1]
[1]Department of English Language & Literature / National & Kapodistrian University of Athens, Greece

Data-Driven Learning (DDL), despite being a feature of corpora and language learning research for some time, is still a relatively unexplored area in the classroom-based assessment literature. Having said that, this presentation reports on an empirical study with young EFL learners (12-15 years old) that aimed at assessing the development of their reading comprehension and language awareness competence through corpus-based activities and data-driven learning situations. The present study followed an experimental approach in order to investigate whether young EFL learners' reading comprehension competence could be improved when exposed to authentic language examples, while setting up situations in which students could answer questions about language themselves by studying corpus data in the form of concordance lines and extended sentences. A total of 60 young EFL students took part in the study and data analysis of pre- and post-achievement tests showed significant improvement in participants' overall ability to

deduce meaning from context while highly competent learners also started using some idiomatic expressions in their written scripts. The findings of the study could provide practical guidance to EFL instructors, materials developers and test-designers as regards the beneficial effect corpus-based activities and data-driven learning can have on young EFL learners' overall language competence.

9:30 - 10:00 Assessment Literacy – How does being an examiner enhance teachers' understanding of assessment?

Victoria Coleman[1], Martin Johnson[1]
[1]Cambridge Assessment, United Kingdom

Concerns have been raised that many teachers do not have sufficient Assessment Literacy (AL), and this has implications for teacher professionalism and classroom practice. AL is an important component of teacher professionalism. It encompasses the basic understandings, skills, and applications that underpin a teacher's ability to use and understand assessment. AL also encompasses a teacher's beliefs and feelings about assessment. This means that the relationship between AL and assessment practice is complex and multidirectional.
Thinking about the transformation of teachers' AL, it is useful to use the metaphor of an 'assessment career'. AL is changeable over time and is influenced by both personal and professional experience. This makes it of interest to explore whether and how teachers' participation in formal examining influences their AL.
To explore the influence of examining on their AL we used concept maps and interviews with a sample of Science and English teacher-examiners. These were then used to develop a survey to explore the influence of examining on the development of AL amongst a wider sample of international teacher-examiners. The outcomes of our study will investigate the contribution that professional examining work has on transforming teacher's AL and any impact on their teaching practices.

10:00 - 10:30 Understanding educators' classroom assessment needs: Using human- centered design principles to establish an assessment use case

Leanne Ketterlin Geller[1], Tina Barton[1], Lindsey Perry[1]
[1]Southern Methodist University, United States

Results from classroom assessments are intended to inform teachers' instructional practices. As the end-users of classroom assessment data, teachers' decision-making needs, preferences, and prior experiences should be considered when specifying a use case (e.g., purpose and intended uses and interpretations). However, there is often a disconnect between the espoused and actual uses of classroom assessment data. To address this divide, principles of human-centered design (HCD) can be integrated into the test development process. HCD seeks to better understand end users' perceptions and perspectives that can serve as the basis for defining the problem and initializing the ideation phase of a solution framework from a user-centered perspective.
In this presentation, we describe the HCD process and outcomes of the iterative development of the use case scenarios and accompanying score reports for classroom assessment resources in early mathematics. Data were collected through a series of structured focus groups with eight elementary school teachers. We share findings from these focus groups and present the prototypical use case scenarios and score reports that were grounded in teachers' decision-making needs, preferences, and prior experiences. We discuss implications for authentic test development practices that are reflective of the end users' needs.

Session EE: Papers 96-98 – National Tests and Examinations II
Chair: Rose Clesham, Room: Castelo 10

9:00 - 9:30 Age-standardising on-demand tests: Is there an effect of "learning time"?
Ben Smith[1]
[1]AlphaPlus, United Kingdom

AlphaPlus is currently leading a consortium to develop a new suite of national assessments for learners aged 7-14 in Wales. These cover three content areas: reading; procedural numeracy; and numerical reasoning, and are administered in both Welsh and English. The consortium is developing and implementing the computer adaptive assessments, with Cito providing expertise and support on the adaptive algorithm.
The first CAT (procedural numeracy) went live in 2018, and as of summer 2019 a full national cohort has completed this assessment. Reporting on learner performance in these assessments has a formative focus, and standardised scores are provided to teachers as an indication of learners' performance relative to the rest of their cohort. The "age effect", wherein older learners tend to score higher than learners who are younger (within their year group), is well-known, so age-standardised scores are also provided.
However, the on-demand nature of the new adaptive assessments introduces a further dimension; learning time (or the time within the academic year at which the test was sat). This presentation reports on our investigation of whether there is a "learning time" effect on performance, and outlines how this can be accounted for when computing standardised scores.

9:30 - 10:00 Exploiting the longitudinal data from exhaustive assessments to measure skills and progress during the first years of schooling
Marianne Fabre[1], Thomas Portelli-Tronville[1], Léa Chabanon[1]
[1]Direction de l'évaluation, de la prospective et de la performance [DEPP], France

In France, an exhaustive assessment, repeated three times over the three first semesters of primary school, has been implemented since September 2018. Nearly 800 000 students are concerned. Called Repères CP-CE1, it fits in with the "Response to Intervention" strategy, a pedagogical approach for the early detection of low performer children.
Besides, it provides an exhaustive and longitudinal data set to estimate progress predictors and process. The size of the data set allows to model heterogeneity and the nested structure of the school system properly.
Analyses combine several types of modeling: regression can predict and explain academic improvement and success, including heterogeneity of outcome and process. Relevant groups are determined by a preliminary clustering. Repeated measurements enable to estimate autoregressive models to study if the rank of children changes over time, and differences in improvement between and within students is explored through hierarchical linear modeling.
Test and estimate of causal relationships between observed and unobserved variables are made through the estimation of Structural Equation Models (SEM). This approach is more general and flexible than regression, and growth of mastery and success modeling factors can be measured by multiple indicators, thus reducing measurement error.

10:00 - 10:30 Predicting grades in external summative assessment of graduates: example from Nazarbayev Intellectual Schools

Daulet Shadiyev[1], Zamira Rakhymbayeva[1], Yerbol Almenov[1]
[1]Nazarbayev Intellectual Schools, Kazakhstan

Successful passing of final grade examinations is decisive step in entering university. The first step to increase university admission chances is the identification of at-risk students early in academic year. Machine learning techniques can provide invaluable insights in predicting students grades on all subjects and allows to inform both teachers and students to make timely decisions.
Nazarbayev Intellectual schools in collaboration with Institute of Pedagogical measurements, Netherlands (CITO) and Cambridge Assessment International Examinations, the UK (CAIE) developed monitoring system of students' progress and system of criteria-based assessment, which includes external summative assessment of students' achievements
The goal of the research is to predict the students' final examination grades using such attributes as summative assessment marks from grade 10, level performance from monitoring mathematics, school, age and gender. Supervised learning algorithms were implemented using Rapid Miner to create prediction model. The dataset contains training and testing data of 2156 2018 graduates, and 2039 2019 graduates as predictive data. After 2019 graduates receive examination grades, accuracy is recalculated, and dataset is enriched.
The research aims to determine preprocessing technique and the model which gives the highest accuracy.

Session EEE: Papers 99-101 – National Tests and Educational Change
Chair: Sandra Johnson, Room: Castelo 3

9:00 - 9:30 External evaluation as a tool for school development: how do Flemish teachers and school leaders engage with school-level feedback from large-scale national assessments?

Evelyn Goffin[1, 2], Mieke Heyvaert[2], Isabel Laenen[2], Rianne Janssen[2], Jan Vanhoof[1]
[1]University of Antwerp, Belgium
[2]KU Leuven, Belgium

In Flanders (Belgium), primary and secondary schools who participate in a large-scale national assessment receive a confidential feedback report about their performance. These reports provide criterion- and norm-based school level feedback. They are a unique monitoring tool in the Flemish educational context because they provide schools with benchmarking information and output data that they can use to inform their practice and policy.
Empirical research on data use indicates that Flemish schools make only limited use of standardized data and feedback from scientific studies. However, no prior research has focused specifically on schools' use of national assessment feedback for school development. As a starting point in an extensive user study, we designed and administered a questionnaire on schools' understanding and use of assessment feedback. In an endeavor to identify promoting and impeding factors, we employed the framework of the Theory of Planned Behavior to include measures for attitude, perceived expectations, self-efficacy and support.
In the presentation we will elaborate on the rationale behind the study and present and discuss some salient results from the questionnaire. We will discuss the implications of our findings for practice, policy, and research on data use.

9:30 - 10:00 Census evaluation as a tool to support educational change: the case of Science education in Peru

Yoni Arámbulo Mogollón[1], Carmen Maribel Carpio[2], Caroline Jongkamp[3]
[1]UMC, Oficina de Medición de la Calidad de los Aprendizajes, Peru
[2]UCR, Universidad Nacional de Costa Rica, Costa Rica
[3]Cito, Institute for Educational Measurement, Netherlands

The national census evaluation of students (ECE) is a large scale evaluation applied annually to collect information on learning achievement of students in basic education in Peru. The ECE has been applied since 2007 for the areas of communication and mathematics for 2nd grade primary students (age 7 years). Since 2016, the census evaluation ECE is being extended to more areas and more grades of evaluation.
The paper discusses the development and first implementation of the ECE for the area of Natural Sciences and Technology for 2nd grade secondary education (age 13 years). The presenters will share the results of the first national census evaluation for Science in 2018 and the approach to promote the effectiveness of learning and scientific thought by means of this evaluation. They will discuss the implications it may have on educational policy on the national, regional and local level, as well as the influence on teaching and learning at the level of the school, teacher and students.

10:00 - 10:30 Implementation of a Pupil Monitoring System on Curaçao to enhance learning outcomes

Wil Knappers[1], Esther Padilla-Bomberg[2], Frans Kleintjes[1], Servaas Frissen[1]
[1]Cito, Netherlands
[2]Curaçao Expertise center for Tests & Exams, Netherlands Antilles

In an effort to bridge the gap between intended and achieved curriculum the Curaçao Expertise center for Tests & Exams (ETE), explored in 2009 possibilities of introducing a pupil monitoring system with all stakeholders. It was expected that by monitoring the growth from an early stage in the students career, using this formative type of the assessment, teachers will get feedback regarding the achievement of pupils to adapt their teaching.
Cito, based on its experience in developing student monitoring systems, supported ETE in the development of a system consisting of a battery of items suitable to monitor the progress against the learning objectives delineated from the intended curriculum. The system for mathematics, covering grades 3 to 7, was developed first and is operational since school year 2017/18. Development of a system to monitor Dutch and Papiamentu is also on the way.
One of the major goals of a pupil monitoring system is to report growth over time, using item response modeling. In addition a more formative way of reporting has been developed. The core consists of reporting on a student's achievement on rather detailed learning objectives. A 'traffic light' reporting indicates whether a pupil has achieved a certain objective.

10.30 - 11.00 Coffee break

11.00 - 11.45 Keynote speech

Chair: Rolf V. Olsen, Room: Castelo 1-2

Title: Improving student's performance with Active Learning

Prof. Xavier Gimenéz

Abstract

Is it possible to practically implement active learning techniques, so that students really improve their learning performance, at the university level? The answer is clearly positive, as evidenced since long by primary and secondary schools, and more recently by engaging initiatives in higher education, mainly in Europe and North America.

The talk will review successful practical implementations of active learning at the classroom level, focusing on the necessary changes that organizations, classrooms and teachers must address in their everyday practice.[1] Supporting evidences will be provided from the overwhelming scientific literature nowadays available. Such evidences have been organized under a "frequently asked questions" format, as it has proven adequate when addressing hesitating stakeholders.

The speech will also present the recent conclusions of the Thematic Peer Group on Promoting Active Learning and Teaching in Universities, more specifically from the paper issued by the European University Association in 2019.[2] The main aspects are: a) implementation of active learning has to be done concurrently at all institutional levels; b) students must have a major role in driving and assessing such implementation; c) teacher career paths must coherently valorize the teaching activity; and d) evidence–based learning and teaching requires an adequate communication strategy, so as to foster the necessary change in mindset.

Short bio

Xavier Giménez Font (Barcelona, 1963) is Professor of Chemistry at the Chemistry Department of the University of Barcelona. He currently teaches Environmental Chemistry and Physical Chemistry of Materials, researches in Computational Simulation of Molecular Systems, speaks and writes widely about popular science and, last but not least, he is much involved in teaching innovation. He did research stages at the University of Perugia, Italy, CNRS in Paris, as well as the University of California, Berkeley. He is author of more 100 research papers and four books about popular science and the teaching of chemistry. He belongs to the Active Learning and Teaching Thematic Peer Group of the European University Association, having created the synchronous flipped classroom methodology SABER, that is used in several Universities as one of their active learning methodologies.

11.50 - 12.35 Keynote speech

Chair: Christina Wikström, Room: Castelo 1-2

Title: Using Response Process Data for informing Group Comparisons

Prof. Kadriye Ercikan

Abstract

Group comparisons, such as gender, ethnic, cultural groups or low or high performing students, are one of the key uses of assessment results. One goal of comparing groups is to gain insights to inform policy and practice and the other is for examining the comparability of scores and score meaning for the comparison groups. Such comparisons typically focus on examinees' final answers and responses to test questions. In this presentation my goal is to discuss and demonstrate the use of response process data in enhacing methodologies used in comparing groups. Response processes may reveal important information about differences in engagement of students that may not be captured by the final responses and provide insights about differences in response patterns that may be identified by using final responses. I argue for use of response process data in addition to final responses to test questions in comparing groups and for examining measurement comparability. I demonstrate use of process data in comparing groups in three example cases. In Study 1, I examine response times for English

Learners (EL) and Non-EL groups on a mathematics assessment and explore how such differences may inform measurement and measurement comparability. In Study 2, I examine sequences of actions captured in key stroke data for EL and Non-EL students on a writing assessment. In Study 3, I focus on using response times in examining measurement comparability in an international assessment. Across these examples I discuss distinctions between response process differences that may constitute measurement inequivalence and others reflecting group differences in engagement with the test which do not constitute measurement inequivalence.

Short bio
Kadriye Ercikan is the Vice President of Psychometrics, Statistics and Data Sciences at ETS and Professor of Education, at the University of British Columbia. She is the current Vice President of American Educational Research Association, a member of the AERA Executive Board of Directors, a member of the ITC Executive Council, has been a member of NCME Board of Directors. She is the recipient of the Significant Contributions Award from AERA Division D.

12.35 - 13.00 Awards and Closing Session
Chair: Jannette Elwood, Room: Castelo 1-2

13.00 - 14.00 Lunch

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://www.aea-europe.net/

AEA-Europe | The Council

President | Jannette Elwood
Queen's University, Belfast, United Kingdom
j.elwood@qub.ac.uk

Vice President | Christina Wikström
Department of Applied Educational Science/Educational Measurement, Umeå University, Sweden
christina.wikstrom@umu.se

Executive Secretary | Alex Scharaschkin
AQA, United Kingdom
AScharaschkin@aqa.org.uk

Treasurer | Cor Sluijter
Cito, Institute for Educational Measurement, Netherlands
cor.sluijter@cito.nl

Council member | Andrej Novik
SCIO, Czech Republic
anovik@scio.cz

Council member | Rolf V. Olsen
Centre for Educational Measurement (CEMO), University of Oslo, Norway
r.v.olsen@cemo.uio.no

Council member | Deborah Chetcuti, University of Malta, Malta
deborah.chetcuti@um.edu.mt

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 2018 committee members are:

- Gill Stewart, SQA, Chair (resigned July 2019) (United Kingdom)
- Amina Afif, Luxembourg Government (Luxembourg)
- Deborah Chetcuti, University of Malta (Malta)
- Mary Richardson, UCL Institute of Education (United Kingdom)
- Lesley Wiseman, University of Glasgow (United Kingdom)

AEA-Europe | Professional Development Committee

The broad objective of the AEA-Europe 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. From 2018 committee members are:

- Rolf V. Olsen, Chair (Centre for Educational Measurement (CEMO), University of Oslo, Norway)
- Stéphanie Berger (University of Zurich, Switzerland)
- Andrew Boyle (AlphaPlus Consultancy, United Kingdom)
- Bas Hemker (Cito, Netherlands)
- Elena Papanastasiou (University of Cyprus, Cyprus)

AEA-Europe | Audit Committee

- Graham Hudson (GA Partnership, United Kingdom)
- Fazilat Siddiq (Nordic Institute for Studies in Innovation, Research and Education, Norway)
- Sebastiaan de Klerk (Ex:plain, Netherlands)

AEA Europe | Conference Local Organising Committee

- Amália Costa (IAVE, Instituto de Avaliação Educativa, Portugal)
- Margarida Borges (IAVE, Instituto de Avaliação Educativa, Portugal)
- Natália Nunes (IAVE, Instituto de Avaliação Educativa, Portugal)

AEA Europe | Conference Scientific Programme Committee

- Co-Chair: Stuart Shaw (Cambridge Assessment, United Kingdom)
- Co-Chair: Andrej Novik (SCIO, Czech Republic)
- Nico Dieteren (Cito, Netherlands)
- Elisa de Padua (University of Cambridge, United Kingdom)
- Pedro Guilherme Rocha dos Reis (University of Lisbon, Portugal)
- Gerry Shiel (ERC, Ireland)

AEA Europe | Review Panel

The council is very grateful for the contribution of all members of the review panel:

- Andrew Boyle (AlphaPlus Consultancy, United Kingdom)
- Angela Verschoor (Cito, Netherlands)
- Anton Béguin (Cito, Netherlands)
- Ayesha Ahmed (University of Cambridge, United Kingdom)
- Cor Sluijter (Cito, Netherlands)
- George MacBride (University of Glasgow, United Kingdom)
- Guri A. Nortvedt (University of Oslo, Norway)
- Iasonas Lamprianou (University of Cyprus, Cyprus)
- Newman Burdett (Chartered Institute of Educational Assessors, United Kingdom)
- Paul Newton (Ofqual, United Kingdom)
- Rose Clesham (Pearson UK, United Kingdom)
- Sarah Maughan (AlphaPlus , United Kingdom)
- Sandra Johnson (Assessment Europe, France)
- Thierry Rocher (Directorate for Assessment, Forecasting and Performance (DEPP), France)
- Wil Knappers (Cito, Netherlands)
- Alex Scharaschkin (AQA, United Kingdom)
- Dina Tsagari (Oslo Metropolitan University, Norway)
- Jana Straková (Institute for Development and Research in Education, Czech Republic)
- Bas Hemker (Cito, Netherlands)
- Rolf V. Olsen (CEMO, University of Oslo, Norway)
- Saskia Wools (Cito, Netherlands)
- Tandi Clausenmay (Learning Works, United Kingdom)
- Rianne Janssen (KU Leuven, Belgium)
- Jasper Wouda (Cito, Netherlands)
- Hendrik Straat (Cito, Netherlands)
- Eva de Schipper (Cito, Netherlands)
- Jannette Elwood (Queen's University Belfast, United Kingdom)
- Remco Feskens (Cito, Netherlands)
- Jude Cosgrove (National University of Ireland, Ireland)
- Anna Lind Pantzare (Umeå University, Sweden)
- Hanna Eklöf (Umeå University, Sweden)
- Carla Pastorino (University of Cambridge, United Kingdom)
- Christina Wikström (Umeå University, Sweden)
- Hana Vonkova (Charles University, Czech Republic)
- Frans Kamphuis (Cito, Netherlands)
- Hasan Selcuk (Charles University, Czech Republic)
- Stuart Shaw (Cambridge Assessment, United Kingdom)
- Andrej Novik (SCIO, Czech Republic)
- Elisa De Padua (University of Cambridge, United Kingdom)
- Gerry Shiel (Educational Research Centre, Ireland)
- Filomena Araujo (IAVE, Portugal)
- Pedro Reis (University of Lisbon, Portugal)

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 2019 panel consisted of three senior assessment researchers. 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 2019, the review panel were Anton Béguin (Netherlands), Rose Clesham (United Kingdom) and Elena Papanastasiou (Cyprus)

The 2019 Kathleen Tattsersall New Researcher Award Winner is Dr. Aisling Keane.

AEA-Europe | The Accreditation review panel

The council is very grateful for the contribution of members reviewed accreditation applications:

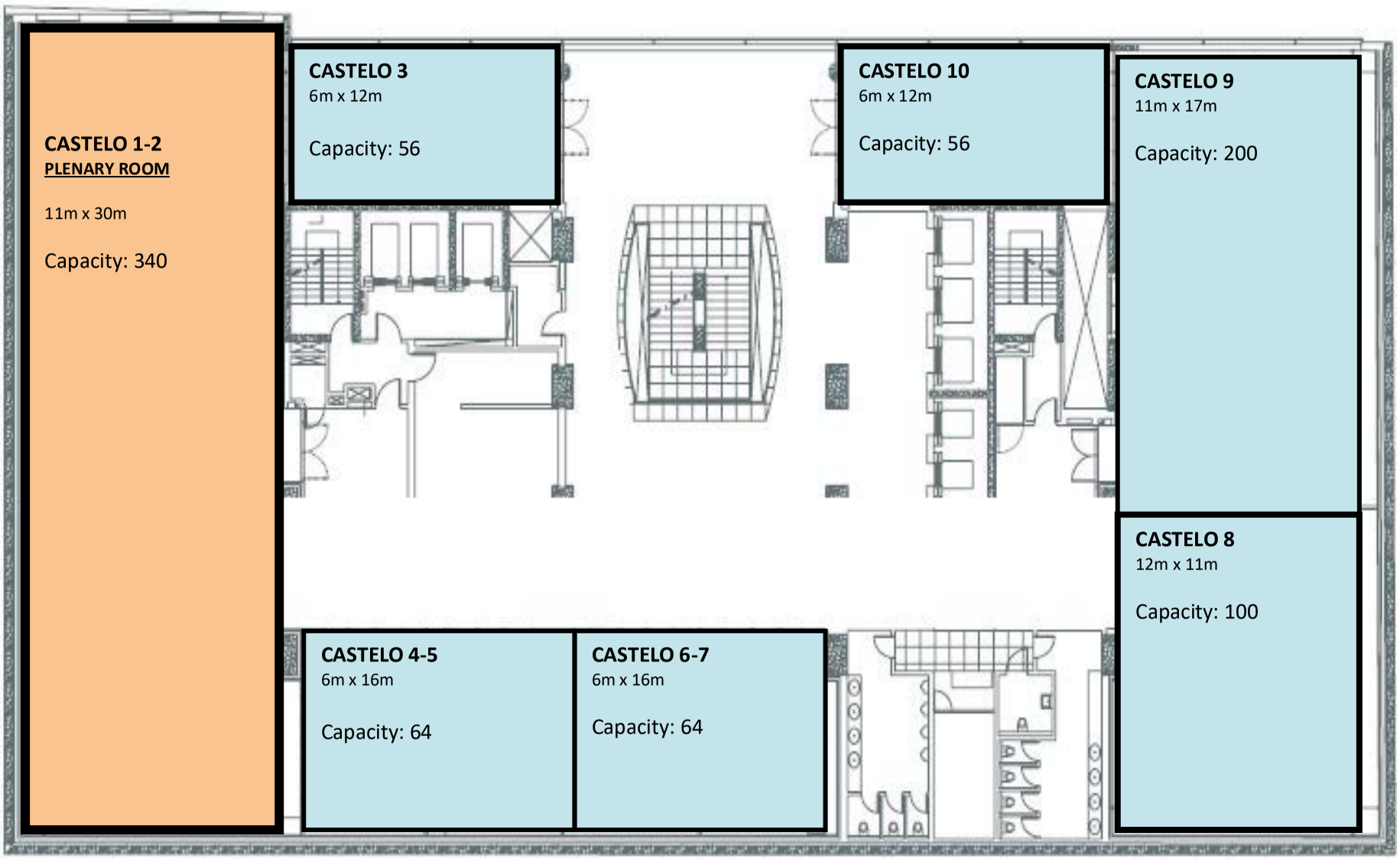
- Ayesha Ahmed (United Kingdom)
- Andrew Boyle (United Kingdom)
- Newman Burdett (United Kingdom)
- Stephen Dobson (Australia)
- Mark Dowling (United Kingdom)
- Jannette Elwood (United Kingdom)
- Bas Hemker (Netherlands)
- Therese N. Hopfenbeck (United Kingdom)
- Ya Ping Hsiao (Netherlands)
- Frans Kleintjes (Netherlands)
- Vasu Krishnaswamy (United Kingdom)
- Sarah Maughan (United Kingdom)
- Paul Newton (United Kingdom)
- Rolf Vegar Olsen (Norway)
- Elena Papanastasiou (Cyprus)
- Cor Sluijter (Netherlands)
- Angela Verschoor (Netherlands)
- Christina Wikström (Sweden)
- Saskia Wools (Netherlands)

Notes

Notes


SANA LISBOA HOTEL

Sana Lisboa Hotel - Rooms



Sponsors

AQA Realising potential

Cambridge Assessment

CITO

GA Partnership

GradeMaker

IAVE Instituto de Avaliação Educativa, I.P.

TAP Air Portugal — A Star Alliance Member

Vretta

AEA-Europe | Association for Educational Assessment - Europe

Assessment for transformation Teaching, learning and improving educational outcomes

The 20th Annual AEA-Europe Conference