AEA-Europe | Association for Educational Assessment - Europe

Assessment cultures in a globalised world

The 18th Annual AEA-Europe Conference

Programme | 8-11 November 2017

Prague, Czech Republic

aea EUROPE

SCIO

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

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

Vice President | Jannette Elwood
Queen's University, Belfast, United Kingdom

Executive Secretary | Alex Scharaschkin
AQA, United Kingdom

Treasurer | Cor Sluijter
Cito, The Netherlands

Contents

Introduction

AEA-Europe is back once more in Prague!

For its 18th conference, AEA-Europe returns to the beautiful capital of the Czech Republic, which in 2000 hosted the association's inaugural conference.

Since 2000, Europe has undergone much change at the institutional, economic and cultural levels. Europe has also faced several crises (financial, migration, identity, etc.). Hence the landscape of 2017 is certainly not one which Europeans in 2000 could have imagined.

One thing which has remained steady, however, and is beyond institutional and political challenges is the growth in exchanges, which increase yearly particularly in the world of education and research. AEA-Europe contributes to this growth: over the years the association has continuously grown and continues to promote activities and trends in the field of assessment. This has fostered networking and collaboration throughout Europe and beyond.

Given this history, the conference theme of 2017 "Assessment cultures in a globalised world" is indeed relevant and timely. The return to Prague will also be the time to take stock through the question of how to reconcile globalisation and identity, community and singularity, local interests and global stakes.

This theme has clearly provoked interest: again, the number of conference submissions has increased, reflecting the attractiveness and drive of the association. With regards to the conference format, as in previous years, in addition to the pre-conference workshops, the conference programme consists of keynotes, open paper sessions, discussion groups and poster presentations. We have retained the innovations introduced last year, such as the poster session with short oral presentations. We have also proposed a new format related to the symposium, where the best-rated one will be entitled to a keynote symposium.

It is important to note that this conference would not take place without the work and commitment of many actors. On behalf of the Association's Council, I would like to convey my heartfelt thanks to the Czech team from Scio who welcomes us here in Prague and is doing everything possible so that the conference is a success. I would also like to thank the members of the Organising Committee, the Programme Committee, the company's name is Easy Conferences and the sponsors who managed the organisation of the conference. Finally, I would like to thank all the people committed in one way or the other to the activities of the Association, especially the members of the Publications Committee and the Professional Development Committee, and those who reviewed the proposals and who have agreed to chair the sessions. It is thanks to the dedication, commitment and willingness of all of these people that this conference will be a success.

In 2000, the conference theme was "Improving Assessment in Europe". Today, we can undoubtedly say that the Association has very much largely contributed to improving educational assessment and that it will continue to do so over time.

AEA-Europe President

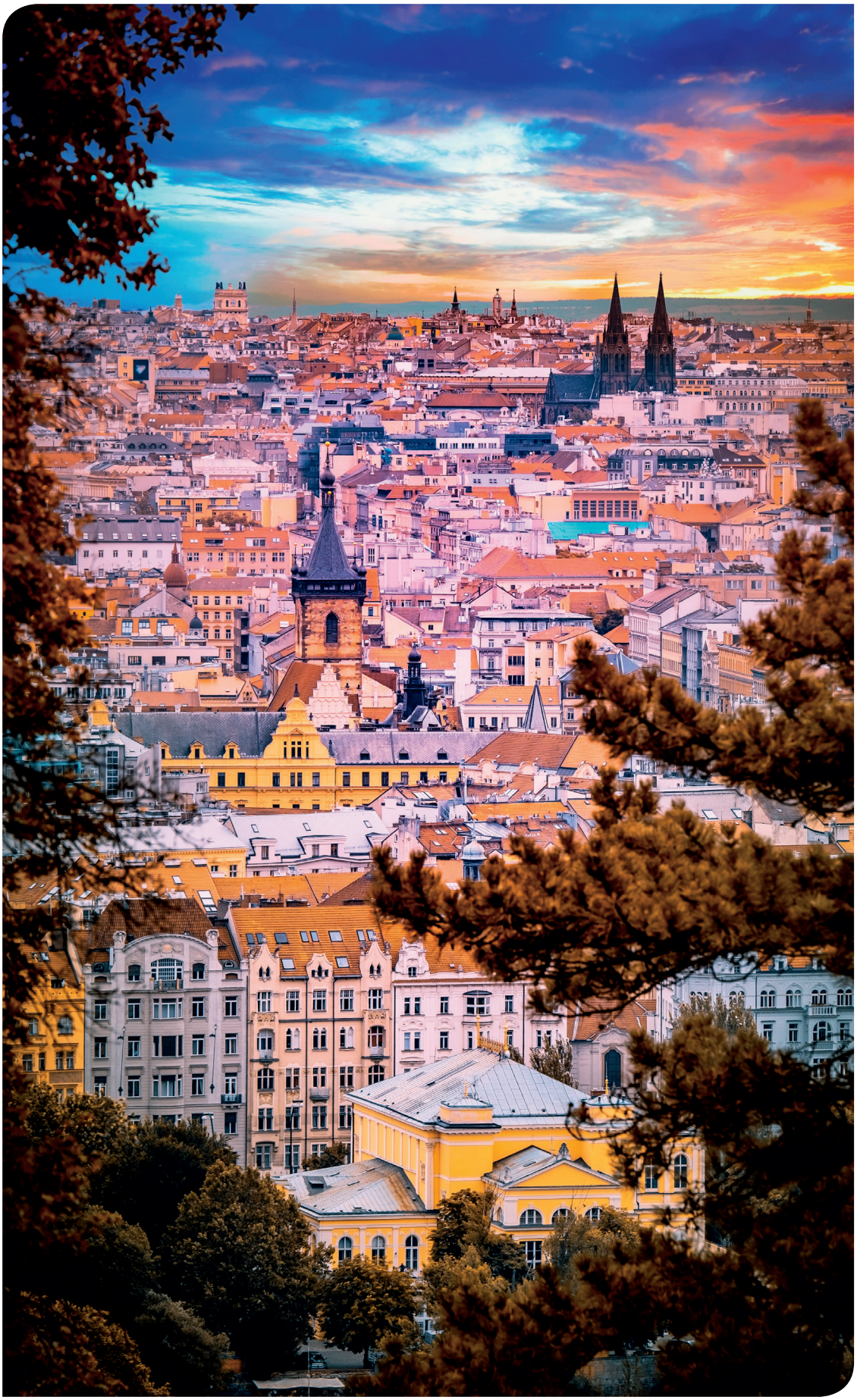
Theme: Assessment cultures in a globalised world

Assessment is a complex and multifaceted practice. Yet many assume that when we talk about 'assessment' we have a common understanding of what it is and what it involves; that there is a universal understanding of assessment. However, this, as we know from experience, is not the case. For example, teacher assessment in some countries is afforded much more authority, is inherently trusted and is seen as professionally sound in much the same way as national examinations are. In other countries, teacher assessment is afforded less power and is considered less reliable than external assessments. Indeed, even examinations tend to have different status and purposes across various educational systems.

Within Europe, different assessment cultures can be observed. In the Scandinavian countries, for example, teacher assessment dominates, while in England, although teacher assessment has been around for a long time, it is less well-regarded by policy makers than external examinations which have been historically provided by a variety of awarding organisations. In other jurisdictions, for example in the countries of Eastern Europe, national examinations are highly trusted and ministries of education are usually responsible for one external examination which all students complete. In all our nations, assessment has taken on what can be defined as a 'cultural script', i.e. a nation can have a propensity to assess in a particular way that is aligned to cultural beliefs about what assessment is and how it should be conducted. For example, in some countries (e.g. Netherlands) standardized testing has a long history and is deeply rooted in the school assessment systems. However, such practices may be less common and not as widespread within the education systems of other countries (e.g. France), which have historically looked to more open-ended ways of assessing students' knowledge. It is interesting to reflect, in a changing European landscape, whether we can truly identify a European tradition of assessment or whether the differences are more striking than the similarities. As a European association, we also reside within a more international, globalized educational arena and it may be that here too traditions and practices of assessment across cultures, while located firmly within national preferences, are becoming more alike because of a more globalized education project and employment market place. Most developed countries participate in one or more of the international comparative assessments such as PIRLS, TIMSS, PISA and ICCS. What is the true value of such assessment practices? Are the results of these international assessment systems overly influencing policymaking? Or, is it inevitable that there is adaptation and mutation of our assessment systems along with other social and political changes more generally?

Even locally, within countries, different stakeholders might well understand assessments differently. Parents, students, politicians and the media sometimes conceive of examinations as absolute and objective while assessment developers are concerned about levels of uncertainty or measurement error. As more and more national tests are introduced into national education systems, communicating about test outcomes and how they might inform teaching, learning and/or policymaking has become a major challenge for test developers, researchers and academics. National standards and school accountability concerns, while of intrinsic value to particular audiences, might well counteract the positive backwash on teaching and learning of good assessment practice and principles. How do those responsible for assessment development and regulation meaningfully interact with key stakeholders (e.g. students, parents, teacher, principals and school governors) in discussing and debating such dilemmas?

Furthermore, how do assessment practices and systems take into consideration the movement and integration of people across and within countries and allow for the variety of experiences and understandings of education and its purpose that different groups bring to the assessment arena. Can we develop assessments and assessment systems that are culturally sensitive and allow every student to show us what they can do?



Programme

Wednesday, 08th November

9.00 – 9.30 Coffee and registration

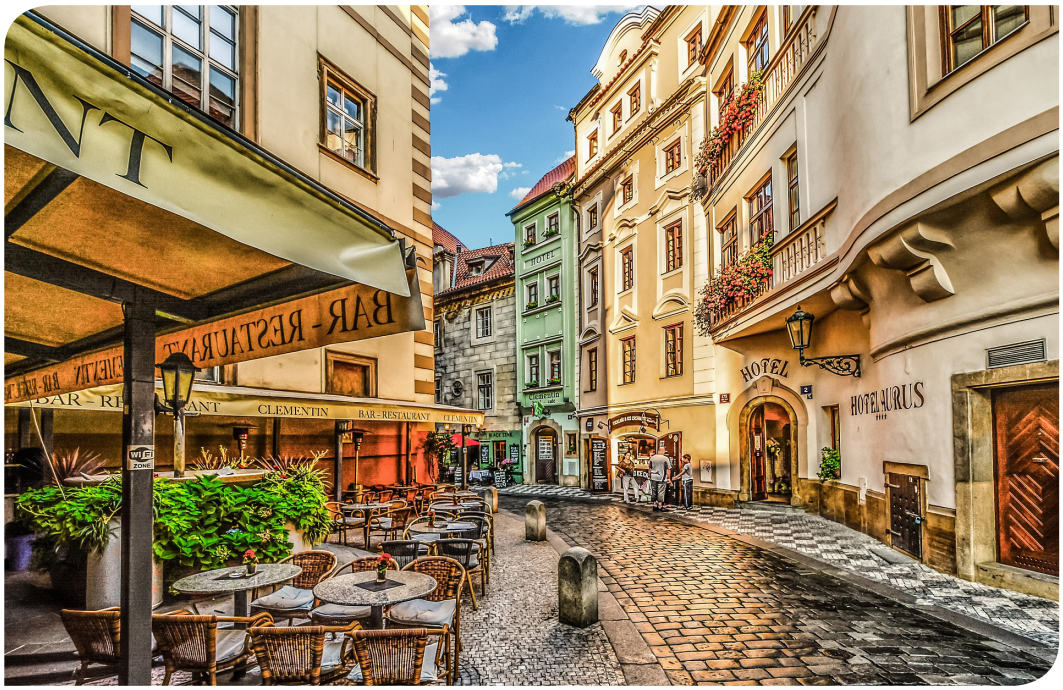
9.30 – 16.30 The pre-conference workshops take place in the CEVRO institute. The address is Jungmannova street 28/17, Prague 1. Nearest Metro stations are Můstek (B, A lines), and Muzeum (C, A lines).

Workshop room 1

9.30 – 16.30 **Applying Test Score Equating Methods using R**

Presenters: *M. Wiberg[1], J. González[2]*
[1]Umeå University, Sweden
[2]Pontificia Universidad Católica de Chile, Chile

The aim of test equating is to adjust the test scores on different test forms so that they can be comparable and used interchangeably. This is extremely important in order to provide fair assessments to all test takers. The goals of the pre-conference workshop are for attendees to be able to understand the principles of equating, to conduct equating, and to interpret the results of equating in reasonable ways. Different R packages will be used to illustrate how to perform equating when test scores data are collected under different data collection designs. Traditional equating methods, kernel equating methods, item response theory (IRT) equating methods and local equating methods will be illustrated. The main part of the training session is devoted to practical exercises in how to prepare and analyze test score data using different data collection designs and different equating methods. Recent developments in equating are also discussed and examples are provided. Expected audience includes researchers, graduate students and practitioners. An introductory statistical background as well as experience in R is recommended but not required.



Workshop room 2

9.30 – 16.30 Item banking for optimal tests

Presenters: *A. Verschoor[1], C. Jongkamp[1]*
[1]Cito, Netherlands

The workshop will offer an introduction into Item Banking and applications for test assembly from a practical point of view. Participants will gain insight in the do's and don'ts when using an item bank for the purpose of developing assessment instruments, and will receive practical guidelines to use metadata and psychometric theory to assemble optimal tests based on an existing item bank. Participants will have hands-on experience in using automated tools to make linear or adaptive tests, based on Item Response Theory (IRT) or Classical Test Theory (CTT).
Main features of these applications will be addressed in the workshop. Participants will be able to understand and assess the usefulness of item banking in their own work.

Workshop room 3

9.30 – 16.30 Comparative Judgement for Research and Practice: an Application of D-PAC

Presenters: *S. Verhavert[1], S. De Maeyer[1], R. Bouwer[1], T. van Daal[1]*
[1]University of Antwerp, Belgium

By the end of this workshop the participants will be familiar with the basic principles and techniques behind assessment with Comparative Judgement (CJ). We will provide a theoretical introduction into CJ and give insight into the steps and decisions needed to set up a study or an assessment using CJ. This workshop will contain two parts. In the morning sessions the participants will get to know the basics behind CJ in an interactive way. Through discussion and practical exercises participants will be encouraged to think about how they could set up a CJ assessment for their own educational or research purpose. In the afternoon sessions the participants will practice with the design of an assessment in the D-PAC tool (www.d-pac.be) and the analyses of the data. This workshop is intended for researchers and practitioners who use, or intend to use, CJ in their research or assessment. Some basic knowledge on statistics might be useful (but is not a must). All analyses will be conducted in Jamovi (www.jamovi.org), a graphical user interface built on top of R. Knowledge of R is not required but some basic notions might be useful.

Workshop room 4

9.30 – 16.30 Large-scale performance assessments: Problems and potentials

Presenters: *R. Janssen[1], E. Ameel[1], J. De Groof[2], A. Deneire[2]*
[1]KU Leuven, Belgium
[2]University of Antwerp, Belgium

With the recent call for 21st century skills in education, questions arise with respect to the large-scale assessment of these competences. As their essence refers to the students' use of knowledge (Silva, 2009), broadening large-scale assessments from only multiple-choice testing to including performance assessments is deemed necessary. Nevertheless, the use of performance assessments in this context has for a long time been the proverbial elephant in the room (Tucker, 2015). The present workshop wants to look at the problems of performance assessments and to discuss their potential solutions, however, without acknowledging their limitations. Firstly, psychometric issues with respect to the measurement and scoring of performance assessments are explained. Secondly, a framework is presented to evaluate the quality of large-scale performance assessments that focus on quality monitoring at the system level. Thirdly, the design and development of performance assessments are discussed with reference to specific examples of performance assessments from the Flemish national assessment program. Finally, the costs and benefits of introducing large-scale performance assessments are weighed in a discussion with the participants.

Workshop room 5

9.30 – 16.30 **How to develop and design valid, innovative and complex computer-based items? – Discussion, sharing experiences and working with innovative item types in a digital environment**

Presenters: *P. Almarlind[1], M. Abrahamsson[1], P. Åström[1]*
[1]Umeå University, Sweden

Why AEA members should attend this workshop:
The aim of the workshop is to gather people from different countries, contexts and with different perspectives to discuss and share knowledge, experiences and new ideas concerning developing and designing innovative and complex computer-based items in large-scale assessments. The workshop will give participants the opportunity to collaborate by looking into and reacting to internal processes concerning how to develop innovative digital item types. The resulting responses and critical questions will hopefully advance the internal development processes. We want to offer a day that balances presentations, discussions and practical work. With the help of visual examples we can hopefully discuss opportunities, constraints, challenges and threats when developing and designing innovative and complex computer-based items in a more concrete way.

Who this workshop is for:
The workshop is designed to engage educational professionals e.g. assessment developers, researchers, educators from different countries who are working with and/or have an interest in designing innovative and complex computer-based large-scale assessments and items.

18.30 – 19.30 **Registration *(hotel Corinthia Lobby)***

18.30 – 19.00 **Welcome reception for new attendees *(hotel Corinthia Lobby)***

19.00 – 20.00 **Welcome reception *(hotel Corinthia Lobby)***

Thursday, 09th November

8.30 – 9.00 Registration

9.00 – 9.45 **Welcome addresses *(Suite 1)***
Tomáš Hruda (Education Republic, Czech Republic)
Thierry Rocher, AEA-Europe President

9.45 – 10.30 **Keynote presentation *(Suite 1)***
Chair: Jannette Elwood
Title: Do we compare comparable? A potential solution with the anchoring vignette method.
Hana Voňková Faculty of Education, Charles University in Prague, Czech Republic

10.30 – 11.00 Coffee

11.00 – 11.45 **Keynote presentation *(Suite 1)***
Chair: Alex Scharaschkin
Title: ICT Literacy Assessment: Status, Innovations and Future Directions
Fazilat Siddiq (Kathleen Tattersall New Researcher)

Poster Presentations

11.45 – 13.00 Posters
Suite 1, Cor Sluijter

11.45 – 13.00 Is the General Certificate of Secondary Education (GCSE) in England incongruous in the light of other jurisdictions' approaches to assessment?
G. Elliott[1], N. Rushton[1]
[1]Cambridge Assessment, United Kingdom

Educational policy makers consider their strategies and practice in the light of what other jurisdictions are doing; informed by results from international comparisons, such as the Programme for International Student Assessment (PISA) and the Trends in International Maths and Science Study (TIMSS), and by observation of teaching, learning and assessment practices worldwide. There is much information available to researchers, and it can be difficult to structure such investigations robustly.
The General Certificate of Secondary Education (GCSE) is an internationally recognised assessment taken by most students in England in 8-10 subjects at age 16+. Students also take A level assessments in 3-4 subjects at age 18+. Some other jurisdictions test only at 18+. In view of this, the English education community has questioned whether carrying out such tests in England at age 16+ is incongruous in the light of other jurisdictions' approaches.
In response to the question outlined above, this presentation explores the strategies that Cambridge Assessment has used recently to select comparator jurisdictions systematically and analyse available data sources effectively. The results of the investigations are presented, illustrating that the GCSE is not incongruous, and that England's assessment structure is just one of a wide variety of different approaches taken.

11.45 – 13.00 Differences in citizenship competencies between Grade 8 and Grade 12 in Flanders (Belgium)
M. Vandenbroeck[1], L. Willem[1], E. Ameel[1], R. Janssen[1], E. Claes[1]
[1]KU Leuven, Belgium

In today's society citizenship competencies are considered very important as they allow people to participate in and contribute to the development and well-being of the complex society they live in (Eurydice, 2005). In this study we investigate whether there are significant differences in citizenship competencies (knowledge, skills and attitudes) between Flemish students in Grade 8 and in Grade 12 of secondary education. Using Item Response Theory and multilevel analysis, data from about 3000 14-year-old Flemish students in the International Civic and Citizenship Study (ICCS) 2016 and about 4000 18-year-old students in the Flemish National Assessment Civic and Citizenship Education 2016 are analyzed. The comparison between the two groups is made possible through the use of anchor items, both in the administered knowledge test on the political-juridical society and in the student- and school-questionnaires for assessing citizenship skills and attitudes. We expect citizenship competencies to be higher among students in Grade 12 because they have had more opportunities to learn and have more coherent knowledge structures (Geijsel et al., 2012). Moreover, some societal issues (e.g. voting, paying taxes, being in contact with institutions) are more present in the life of 18-year-olds, which may affect their political involvement.

11.45 – 13.00 Assessment literacy in the multicultural science and mathematics classroom: Working with student-teachers to develop an understanding of fairness and equity.
D. Chetcuti[1], J. Farrugia[1], M. Buhagiar[1], C. Calleja[1], M.M. Musumeci[1]
[1]University of Malta, Malta

An important aspect of assessment literacy is the development of an understanding of the issues of 'fairness' and 'equity'. In science and mathematics classrooms that are becoming more culturally diverse it is also important for student-teacher to develop assessment strategies that enable them to assess the learning of students who come from diverse cultural backgrounds in a fair and equitable way.

In this poster presentation we would like to address this issue and report on the development of a 'two-hour assessment module' for student-teachers following the Masters in Teaching and Learning at the Faculty of Education, University of Malta by academics from two departments within the Faculty. The academics from the Department of Mathematics and Science Education and the Department of Inclusion and Access to Learning came together to develop the programme that would help student-teachers be better able to handle assessment in a fair and equitable manner in the multicultural classroom. The poster will include reflections on the dialogue between the academics and provide a snapshot of the differing views surrounding the issues of fairness and equity. The poster will also give description of the assessment module including practical activities that challenge the students' views about fairness and diversity.

11.45 – 13.00 Who is the smartest in the Czech Republic? The analysis of results of entrance exams to Czech universities
T. Habermann[1], L. Fiřtová[2]
[1]Scio, Czech Republic
[2]University of Economics, Prague, Czech Republic

In the Czech Republic, the proportion of students pursuing tertiary education has been rising and so has the number of universities. However, as university entrance exams have a strong tradition in our country, they are still an important part of most students' academic career. The system of university entrance exams in the Czech Republic can be characterized by a certain duality: while some public universities create their own, typically knowledge-based tests, there are still many that opt for using tests created on a commercial basis. As the largest provider of commercially created tests in the Czech Republic, we have many data indicating various interesting trends. The most widespread test used in the context of Czech entrance exams is the test of General Academic Prerequisites, taken by thousands of students each year, about 15% of whom are of Slovak nationality. In our poster, we are going to show how the test results (namely scores in verbal reasoning and quantitative reasoning) differ by gender, motivation and nationality of the examinees, and what these results may imply. Finally, the poster is going to include detailed information on the system of entrance exams in the Czech Republic for those interested in educational policies.

11.45 – 13.00 Defining an Interoperability Standard on CAT (Computer Adaptive Testing)
M. Molenaar[1]
[1]Open Assessment Technologies, Luxembourg

CAT (Computer Adaptive Testing) has been around for a long time. Unfortunately, implementations of CAT have always been proprietary and by definition not interoperable. In other words: once an adaptive test was developed in a certain assessment platform, it was almost impossible to switch platform without significant additional investment.
To address this, an IMS workgroup of industry leaders was established to formulate a Standard on CAT: a set of best practices to extend current interoperability standards (QTI) to contain all required (psychometric) data and define a common language for adaptive engines.

The general consensus is not to define one generic sequencing engine or an elaborate standard which defines all possible variations in CAT, but to treat an adaptive engine as a "black box" and define a common language to communicate with these engines. This way end-users are assured much needed (content) interoperability, while platform vendors and researchers can continue to innovate.
This poster presentation will present the latest developments in this new interoperability standard and aims to facilitate a discussion amongst attendees on CAT interoperability.

11.45 – 13.00 Re-designing the role of examiner judgement in maintaining standards for UK general qualification examinations

J. Maziarz[1], A. Castle-Herbert[1], S. Denner[1], L. Phillips[1], R. Harry[1]
[1]WJEC, United Kingdom

Despite criticism of its reliability, examiner judgement remains crucial in preserving public trust in the UK examination system. This function has become more important as qualification reform and subsequent shifts in cohort size and profile challenge the reliability of statistical approaches to maintaining standards in GCSEs and A levels. Concerns regarding the veracity of existing methods of utilising judgements remain, however.

The replacement of the Code of Practice with more general regulatory Conditions provides an opportunity to re-evaluate and redefine the role of examiner judgment. As one of the UK's major examination boards, WJEC has conducted multi-stage research to develop methods and procedures to maximise the value of examiners' skills and experience to the awarding process, whilst ensuring fairness and reliability.

Firstly, literature was summarised evaluating current 'state of play' regarding the role of judgment and related methods in setting and maintaining of the examination standards. Psychological perspectives were employed to evaluate the limitations and strengths of existing approaches. Based on this, a 'design thinking' approach was used to create and prototype alternative models. Following a preliminary evaluation, a trial of the most promising methods will be conducted during the summer 2017 awarding series.

11.45 – 13.00 Effects of reference set reliability on the efficiency of two-stage comparative judgement

A. Furlong[1], R. Bouwer[2], S. Verhavert[2], S. De Maeyer[2]
[1]International Baccalaureate, Netherlands
[2]University of Antwerp, Belgium

Comparative Judgement (CJ) has emerged as a valid and reliable alternative to marking. However, there is debate about which strategy for pairing student responses together best combines efficiency with reliability. One possibility is a two-stage approach, whereby a CJ session is run on a subset of responses creating a calibrated rank order, and then subsequent responses are judged against that initial reference set using an adaptive algorithm. Currently, there is little research regarding how reliable the reference set needs to be and how that affects the number of judgements needed to appropriately judge other responses against it.

To investigate this, an initial CJ session of 160 essays, 15 judges and 28 comparisons per essay (scale separation reliability (SSR) of 0.91) was run on the D-PAC platform. Four reference sets were created comprising the same subset of 140 essays but with varying SSR values (0.5, 0.7, 0.8 and 0.9) and the remaining 20 essays were re-judged against each of these reference sets. Results focus on the number of judgements required for each essay to be appropriately re-judged and how the final placements of these 20 essays against each of the four reference sets compare to those derived from the initial session.

11.45 – 13.00 Using assessment data to drive school improvement

D. Haggie[1], M. Mackinlay[1]
[1]GradeMaker Ltd, United Kingdom

This poster presents a project GradeMaker is currently delivering, in association with FFT Education, to set up an education data portal in Guyana. The portal is for primary and secondary schools, inspectors, regional government and the Ministry of Education, and presents contextual analysis of exam results data, student data and data about schools.
During the project we have loaded longitudinal data sets for 3 different exam series, matched the data and, in consultation with the Ministry of education, developed analytical reports for all stakeholders. A trial phase has been completed and further data is being loaded. The system is being used to strengthen educational planning at all levels, and to improve school improvement planning. The service is live and will be rolled out nationally to all schools at the start of 2018.

11.45 – 13.00 Eductional Reform in Four Provinces and Their Assessment Approches

N. Wei[1]
[1]NCCT, China

In Year 2001, Curriculum Reform Guidelines on Basic Education was published by the Ministry of Education in China Mainland, which implied the beginning of the curriculum reform.
The Children's entry ratio to primary school reached 99.9% in China mainland (national statistics, 2012). Based on the fact that every child has the opportunity to go to school, the quality of education has been advocated nationally, which calls for assessment projects.
The present research is going to study the educational reform in four provinces at different economic development levels and their assessment approaches, with the employment of qualitative and quantitative research methods. The results show that students' academic performance differentiate while their recognition on learning keep consistent; the devotion on education of four provinces varies, including the ratio of students and teachers, computer equipped, etc., which correlates to their economic development; teachers and principals have similar and different beliefs on education. However, it is easier for provinces with higher level of economic devleopment to take chellenges, including their participation in important educational projects, in national and international assessment projects and in the new university entrance examination reform, etc..

11.45 – 13.00 Changing ability and comparable outcomes – UK examinations of French, Spanish and German

A. Evans[1], A. Castle-Herbert[1], P. Morgan[1]
[1]WJEC, United Kingdom

Background: In much of the UK a system of "Comparable Outcomes" is used to set predictions for how many candidates should receive different grades during their main exams at 17 (AS level) and 18 (A level). Candidates are split into deciles depending on the mean of their exam results at 16 (GCSEs) and predictions for outcomes are set so that each decile is expected to perform to the same historic standard. In French, Spanish and German the numbers entering AS and A level exams has dropped significantly and this may be affecting comparability.

Methods: In addition to looking at the Mean GCSE of candidates taking foreign language AS and A levels we also looked at how they did in the GCSE that corresponded to the qualification they were taking. Analysis was undertaken to see how this relationship changed between 2010 and 2016.

Results: We found that candidates in lower deciles were more likely to have received top grades (A* or A) in 2016 than they were in 2010. This puts into question the assumption that candidates in these deciles were comparable between the two periods.

11.45 – 13.00 What happens when extended response question papers are no longer divided into items for marking?

D. West[1]
[1]*AQA, United Kingdom*

When test item marks are added together as a total mark, how clearly are the three facets of each measurement, candidate, item and rater, contributing to the test scores?

A large scale live empirical comparison between two systems of on-screen marking is presented. 17 A-level extended response question papers in the UK were switched from item level to on-screen whole script marking between the summers of 2014 and 2015. About 142,000 students sat these question papers in each year.

In whole script marking, candidates' overall scores became more spread out. Internal test consistency, measured by inter-item correlation, rose by nearly 50%. The range of percentage marks across different items within a script fell by 15%.

Is this because raters show bias? The results varied little between raters, whose marking was monitored using occasional seeds or double marking. If bias occurs in whole script marking then this is at the level of scripts, not centres or raters. Correlation of marks with prior attainment of the candidates did not change when whole script replaced item level marking.

Simulation studies and variance decomposition methods will be presented which help us to understand the changes in mark distributions between the two marking systems.

11.45 – 13.00 Introducing self-assessments and selv-evaluations to reach out to a wider population.

I. Radtke[1]
[1]*Skills Norway, Norway*

Although Norway does have a system for the provision of basic skills training for adults, the target groups remain to a large extent unaware of their training possibilities. While Norway as a whole scores higher than average in PIAAC, significant percentages of the population score too low.
The current Norwegian government expressed its intention to establish a national commitment towards adults with poor basic skills by: (1) ensuring that people who receive unemployment benefits, should automatically be offered an assessment of basic skills, and (2) establishing a general right to assessment of basic skills for adults.
Studies by Skills Norway show also that the country lacks adequate assessment tools, especially a simple, user friendly, and accessible screening test which can be used both for self-testing and to help professionals in employment agencies or career centers determine if individuals are in need of basic skills training.
The poster shows the result of the project, which is the creation of userfriendly online screening tools in reading, numeracy and IT-skills that can be used by both individuals and counsellors to determine the need for basic skills training.

11.45 – 13.00 From large-scale national assessment data to didactical research: a textbook case study on standard written algorithms

E. Goffin[1], *W. Van Dooren*[1], *E. Ameel*[1], *R. Janssen*[1]
[1]*KU Leuven, Belgium*

The present paper discusses a mixed model case study that examined the relationship between 12-year-olds' performance in standard written algorithms to the mathematics textbook used in class.

A first, quantitative study consisted of a multilevel regression analysis on data from the 2009 Flemish mathematics assessment at the end of primary education. For several mathematics domains, a link was found between the textbook and pupils' learning outcomes. In a second, qualitative study, two popular textbooks that markedly differed in assessment results for the domain of standard written algorithms were compared. An analysis of contrasts in the textbooks' structure and content showed that the textbook associated with better scores attributes more time and more exercises to standard written algorithms and displays different didactic principles. This case study was exploratory in nature and design. As contrasting cases were purposively sampled, the conclusions may not be generalizable. The study illustrates that, although cross sectional system-level data do not allow for making causal inferences, they can provide reliable stepping stones for other disciplines in educational effectiveness research and ultimately for advancements in instructional practice.

11.45 – 13.00 Refocusing assessment: A framework for a managed transition in assessment culture for schools.

D. Thomas[1], C. Hodgson[1], S. O'Farrell[2], A. Galvin[3]
[1]National Foundation for Educational Research, United Kingdom
[2]Association of School and College Leaders, United Kingdom
[3]The Schools, Students and Teachers Network, United Kingdom

The introduction of the new national curriculum in 2014 in England represented an important shift in assessment culture – removing the centralised system of reportable national curriculum levels. Although this provided schools with an opportunity, to refocus assessment on the needs of learners, it also caused uncertainty for those responsible for developing effective assessment policies. NFER worked with a school leaders' association (ASCL) and a schools network organisation (SSAT) to produce a free resource to guide schools through this cultural transition.

The resource focuses on the assessment of 11-14 year olds, providing a framework of key questions to guide schools systematically through discussions to build more formative approaches to assessment. These questions focus teacher discussions on defining and evidencing subject specific progress. In developing the framework we consulted expert panels of subject heads and subject association representatives, collating the experts' responses to the questions into subject documents to support and challenge discussions in schools.

The framework also supports senior leaders in planning a coherent assessment policy informed by the discussions of their own practitioners and based on the needs of their learners. This resource provides a rational framework guiding schools through an important cultural transition in assessment.

11.45 – 13.00 Elimination scoring versus correction for guessing: A simulation study

Q. Wu[1], T. De Laet[1], R. Janssen[1]
[1]KU Leuven, Belgium

Administering multiple-choice questions with correction for guessing fails take into account partial knowledge and may disadvantage examinees who are risk averse. In order to overcome these disadvantages, elimination scoring has been proposed in which examinees need to eliminate all answer options they think are incorrect. The current study investigates how these two scoring procedures affect response behaviors of examinees who differ not only in ability but also in their attitude toward risk. A two-step model is proposed to simulate the expected answering patterns on multiple-choice questions: (1) probabilities of a correct response to each of the alternatives in a multiple-choice question are modeled using the Rasch model based on ability; (2) the decision making of giving a particular answering pattern is modeled using prospect theory that takes risk aversion into account. The results from the simulation study show that overall ability has a predominant effect on the expected scores, while risk aversion has a decisive

impact on expected answering patterns for examinees with intermediate success probabilities on the items. These examinees benefit more from using elimination scoring.

13.00 – 14.00 Lunch

Open Paper Sessions

14.00 – 15.30 Session A: Establishing and Maintaining Standards 1
Douro and Oder, Mary Richardson

14.00 – 14.30 Grading Severity in Malta's National Examinations
G.J. Zahra[1], D. Pirotta[1]
[1]MATSEC Support Unit, University of Malta, Malta

It is commonly held that an examination body should maintain a standard level of difficulty across different years, tiers, and subjects. Nevertheless, grade setting does depend, to a certain extent, on expert (human) judgement and different studies have suggested that the same standard of difficulty is not maintained across different examination boards and subjects. This study examines the results obtained by candidates in 2016 in Secondary Education Certificate (SEC) 16+ national examinations to assess whether some subjects are more severely graded ('difficult') than others.

This study assumes that there is a general academic ability which influences candidates' grades in all subjects. An estimate of this general academic ability (G4) was calculated by using candidates' raw scores in four rather compulsory subjects. The study then measures and compares mean G4 of candidates in identical grades obtained in different subjects at SEC level. It is assumed that, if subjects are graded with the same severity, the mean G4 obtained by candidates for the same grade in different subjects will not vary at a statistically significant level. Although the results show instances of statistical significance, it seems that most differences are not disturbingly large.

14.30 – 15.00 Investigating the features of levels-based mark schemes associated with consistent marking
B. Black[1], S. Humphries[1]
[1]Ofqual, United Kingdom

In common with many other jurisdictions, the standard qualifications for 16 and 18 year-olds in England – General Certification of School Education (GCSE) and A levels – have examinations which often contain extended response items which are marked using levels-based mark schemes. In the endeavour to produce assessments that measure students' performance in a valid way, exam boards must the best method of scoring responses in order to accurately and consistently. However, extended response examination questions which are marked using levels-based mark schemes tend to be associated with less consistent marking.
Over 2000 levels-based mark schemes from 200 examinations in ten GCSE and A level examinations were coded on over 40 features including those identified by previous literature including structure, content and presentational features.
A multi-level regression model revealed associations between some mark scheme features and marking reliability statistics (derived from live marker monitoring data). A number of the noteworthy observations bring together previously established literature on mark scheme use, and some are novel to this study. All are discussed within the context of mark scheme design, the valid assessment of students' work, and the apparent tensions between reliability and validity.

15.00 – 15.30 Is there a Nordic model in education? A comparison of standard-setting practices in the Nordic countries
S. Blömeke[1]
[1]*UiO/ LEA/ CEMO, Norway*

This paper examines the assessment policies in Denmark, Finland, Norway and Sweden – in particular with respect to standard-setting. Standard-setting includes the definition of proficiency levels and corresponding cut-scores. In Sweden, standard-setting has existed since the 1960s and is meant to support teachers' grading. A national curriculum defines the standards, population-based national assessments in years 3, 6, 9 and upper-secondary provide the data. In Denmark, computer-based adaptive population-based national tests were not introduced before 2006. Standard-setting happens from years 2 and 3 on (horizontally and). In Norway, a national curriculum defines only broad objectives. Population-based national testing happens in years 5 and 8 and results in empirically defined standards. No link to grading is made. In Finland, no population-based testing exists at all. Monitoring happens through sample-based national testing and municipal assessments. These differences in assessment policies and standard-setting weaken the assumption of a "Nordic model" in education. They may be explained by differences in the centrality of educational policy and in achievement in international studies. In Finland and Denmark, long traditions of municipal autonomy exist whereas Sweden and Norway had long traditions of state governance. Denmark and Norway achieved mediocre PISA results whereas Finland and Sweden started rather strong.

14.00 – 15.30 Session B: Teacher Assessment Practices
Danube, Deborah Chetcuti

14.00 – 14.30 Towards effective feedback practices: An investigation into teacher and student perspectives
F. van der Kleij[1]
[1]*Australian Catholic University, Australia*

Quality feedback is one of the most powerful influences on student learning. However, the potential of feedback in helping students learn is generally not realised in classroom practice. Earlier research found a discrepancy between feedback practices as perceived by teachers and by their students. Understanding the intended and perceived meaning and value of feedback messages is a critical first condition for effective feedback uptake for student learning.
The present study, therefore, focuses on investigating differences in feedback perceptions among teachers and students. Gaining insights into when, how, and why students find feedback helpful is necessary to better understand and optimise the feedback process.
Teachers filled out an amended English version of an existing Norwegian-developed self-report survey of feedback practice. Students filled out the student version of the survey, as well as a survey measuring background characteristics, including motivation (including self-efficacy and intrinsic values), self-regulation, and ability levels. The study focused on two core subjects, English and Mathematics, in the Year levels 7-10 in five schools in Australia.
Findings suggest that teachers consider their feedback more favourably than students do and showed considerable variation in students' feedback perceptions. Student feedback perceptions were found to correlate with their background characteristics.

14.30 – 15.00 Testing: development of tools and mechanisms for assessing the level of professional knowledge of teachers
B. Bayekeshova[1]*, T. Lakhtina*[1]
[1]*Nazarbayev Intellectual schools, Kazakhstan*

The relevance of the study has been determined by the need for evaluation mechanisms that have certain specified properties to improve the reliability and validity of measuring instruments.

This problem is important for solving a wide range of issues of selection, placement, assessment of the level of skills of personnel, particularly, of the pedagogical staff.
The purpose of the study is to determine the feasibility of developing test tasks with given properties for assessing the professional knowledge of teachers.
In this research, the assessment objectives are correlated with the objectives of the training (knowledge and understanding of modern approaches to the organization of teaching and learning, resulting from them practical skills). Research methodology: classification, statistical processing of test results and comparative analysis. The main result of the study is the conclusions about how the test tasks of these types function. The results of the research may be of interest to specialists involved in testing personnel, including pedagogical staff, in situations of training, retraining and professional development training courses.
Key words: professional knowledge, qualification examination, testing.

15.00 – 15.30 Assessment capacity building MOOCs' – How can we facilitate school-based and teacher professional development that promotes improved pedagogy, assessment and delivers improved learning outcomes? The case of Norway.
E.W. Hartberg[1], V. Meland[1]
[1]Inland Norway University, Norway

MOOCs sit at the interface between a standardised international form of assessment and empowering the assessment taker to choose not only when they take the assessment but also how MOOCs increasingly might possess group functionality or offer a culturally sensitive pedagogy and individually adaptive learning. In this presentation we consider the potential evidenced in a School based Assessment for Learning MOOC developed and delivered since 2015 by ourselves as a paid consultancy for the Norwegian Directorate of Education and Training. To date the MOOC has been taken by 600 schools and approx. 16 000 teachers. The main objective is to build a school based assessment capacity and a culture to promote student learning. Central to this objective is the view that 'the teacher as a stakeholder in the development of new assessment paradigms'.

In our session we will present the MOOC and the research and evaluation project. Addressing all the issues raised above we will critically reflect upon the importance of MOOCs as a national and global vehicle for developing new and innovative assessment assessment paradigms for individual teachers and shared School cultures in which they work.

14.00 – 15.30 Session C: The Differentiating Effect of Assessment Methods on Various Groups of Testees
Amstel and Volga, Martyn Ware

14.00 – 14.30 Assessing primary school children: Does a child's social and cultural background have a differential impact on their performance across different assessment measures?
S. Stothard[1], G. Copestake[1], L. Copping[1], C. McKenna[1]
[1]University of Durham, United Kingdom

Frequent educational assessment is standard practice in England. For example, during primary school (ages 4-11 years) formal assessments include: baseline check of cognitive skills on school entry, assessing phonics at age 6, and measuring English and maths at age 11. Test results are used for a variety of purposes, including school accountability, identifying additional support needs, checking pupil progress, and selection for academically selective schools. There is an implicit assumption that assessments are fair and unbiased. However, are all assessment formats equally fair and valid for all children? Here we report the results of a study examining the performance of a large cohort of children (N=4707) on formal educational assessments from age 6-11 years. We investigate the impact of demographic

variables (e.g., gender, additional language learning, ethnicity, social deprivation) on test performance, and test for possible interactive effects between demographic variables. Children from disadvantaged homes and children speaking English as an additional language gained significantly lower mean scores than their classmates on all test measures, and were consistently under-represented amongst the highest attainers. Correlations between measures were also weaker for disadvantaged groups. We explore the role assessments might play in maintaining group biases, and discuss implications for educational policy.

14.30 – 15.00 Fairness in the selection to higher education – (how) does the choice of methods for assessing and rank ordering the students matter?
C. Wikstrom[1], M. Wikstrom[1]
[1]Umeå University, Sweden

The purpose of this paper is to study differences in performance on two fundamentally different instruments that are used in the selection to higher education in Sweden; a norm-referenced test, and criterion-referenced grades. This is done by comparing how groups of students are ranked on the basis of their total scores and grades, but also how they perform on the quantitative and verbal section scores of the selection test in relation to national tests in English and mathematics and corresponding subject grades from upper secondary school. The data includes information on test takers who took the selection test in the autumn of 2011 (N=23,214) or spring of 2012 (N=27,075), and is analysed with correlations and regression analysis. The findings show that male test takers perform higher on the test and female higher on the grades, but when studying separate sub-tests with grades and national course tests from isolated subjects, the students seem to be ranked more similarly. An interesting finding is that male test takers with a non-Swedish background seem to be graded more leniently than females with a similar background in mathematics, and the opposite is the case in verbal subjects. Potential causes and implications are discussed.

15.00 – 15.30 Do Examinees with Dyslexia Take Longer to Answer Items on a Test of English as a Second Language than Examinees without Dyslexia?
N. Gafni[1], M. Baumer[1], M. Eitan[1]
[1]National Institute for Testing & Evaluation, Israel

The provision of special test accommodations plays an important role in any discussion of a test's fairness and accessibility to examinees with learning disabilities. The goal of accommodations is to ensure that the test in question measures what it is meant to measure among examinees with disabilities in the same way that it measures those attributes among non-disabled examinees.
This study investigates whether examinees diagnosed with dyslexia used more time to answer items on a test of English as a second language (ESL) than examinees without dyslexia, under conditions in which there was no limit on the time allotted for each test item. Study subjects comprised 27,376 examinees without disabilities and 799 examinees with disabilities who took an adaptive, computerized test of ESL.
We assumed that examinees with dyslexia would need more time to answer items on the test than examinees without dyslexia who had the same level of English proficiency. Among most examinees, our prediction was not supported.
The findings raise the issue of whether the allocation of additional time actually impairs the fairness of the test with regard to examinees without disabilities, who would have received a higher score had they been allotted more time.

14.00 – 15.30 Session D: Assessment Culture and its Impact
Loire and Elbe, Stuart Shaw

14.00 – 14.30 The ability to learn – what does it depend on and can we measure it?
L. Fiřtová[1, 2]
[1]Scio, Czech Republic
[2]University of Economics, Prague, Czech Republic

This paper presents the interim results of the SOLE Box project, whose aim is to reduce early school drop-out rates and improve attitudes to learning. The project explores the concept of Self-organized Learning Environment (SOLE) and its impact on children's ability to learn (learner autonomy), the assessment of which is, despite its rising importance, still an underexplored topic. It involves six partner organizations and approximately 75 children from disadvantaged (mainly Roma) communities in Slovakia, Romania and Kosovo.

Over the course of three years, children are going to explore several boxes containing various tasks and activities (environmental games, tools to make a video...). The main idea behind the project is that letting children explore the boxes on their own, with minimal teacher intervention, should also lead to an increase in their level of learner autonomy.

During our presentation, we will discuss whether it is possible to measure learner autonomy, whether the SOLE approach indeed improves one's learner autonomy and whether there is a relationship between the effectiveness of this approach and demographic characteristics, personality and interests. Our presentation will be based on the data collected insofar using questionnaires and structured interviews.

14.30 – 15.00 Assessment systems as cultural scripts: reflections on assessment processes and practices in a shifting social, cultural and globalised world
J. Elwood[1]
[1]Queen's University Belfast, United Kingdom

This conference is focusing on assessment cultures in a globalised world. As the theme outlines, there are a number of assumptions about what we mean when we talk about assessment. What then might we mean by 'cultures' in relation to assessment. Cultural beliefs or 'scripts' about what assessment is become embedded in expressions of assessments' purposes, uses and practices. This paper will focus on a theoretical exploration of what might be meant by a 'cultural script' in assessment and present emerging debates that consider sociocultural theory as offering a coherent set of considerations to understand assessment as it operates in the global world. This paper will also attend to the challenges for assessment raised by sociocultural theories to formative and summative assessment frameworks and arenas. Often what happens in assessment is that, without exposure to alternatives, or understanding of the ways in which assessment interacts differentially with different groups of students, or understanding the ethical and moral impact of assessment choices, students, teachers, assessment developers, and policy makers draw on cultural legacies about assessment in relation to how it should be done, what constitutes rigorous and valued assessment and how this should be played out within national and international systems.

15.00 – 15.30 Universal Quality Criteria
P.E. Newton[1]
[1]Ofqual, United Kingdom

The Scientific Programme Committee began their discussion of the 2017 conference theme by proposing that we do not share "a common understanding of what [educational assessment] is and what it involves". They asked whether it might be possible to "identify a

European tradition of assessment" but their prior exemplification of different "assessment cultures" seemed strongly to imply that it might not be.
Unfortunately, the absence of a common understanding of educational assessment would seem to preclude the possibility of common quality criteria, which might cast doubt over the very idea of a European Framework of Standards for Educational Assessment.
Messick proposed construct validity as a universal quality criterion; the principal criterion, equally applicable across all kinds of assessment formats. However, prominent scholars subsequently questioned its adequacy for judging alternative assessment formats, e.g. performance assessments (vs. traditional tests). This seems to mirror the Committee's invocation of alternative assessment cultures. The present paper will defend the idea of universal quality criteria, arguing that we are united by an understanding of educational assessment that is essentially universal. It will explore reasons why this might not always seem to be the case, including the very different trade-offs that are made across different assessment traditions.

14.00 – 15.30 Session E: Assessing Hard to Measure Constructs 1
Tiber, Tim Oates

14.00 – 14.30 Varieties of Conceptions of Ethical Competence and the Search for Strategies for Assessment in Ethics Education: A Critical Analysis
O. Franck[1]
[1]University of Gothenburg, Sweden

This presentation highlights some challenges related to assessment in ethics education. More specifically, it starts with an analysis of how ethical competence is evaluated in national tests in Religious Education (RE) in Sweden. Seven conceptions of such a competence are identified in the tests as well as in relevant policy documents, and these conceptions, often implicitly embedded in items and descriptions, are analysed with regard to pedagogical as well as philosophical and ethical considerations. Questions regarding consistency, coherence and reliability are raised, and a critical analysis of the general question on the possibility of "assessing ethical competence" is developed, initially with regard to some perspectives within the virtue and capability approach presented by Amartya Sen and Martha Nussbaum. The presentation ends with some suggestions for developing the approach as well as specific items in the national tests regarding ethics, in order to take into account a more complex interpretation of ethical competence and to make this interpretation transparent to pupils and teachers.

The presentation is based on the author´s chapter in Franck, O. (ed) (2017): Assessment in Ethics Education – A Case of National Tests in Religious Education, Dordrecht: Springer.

14.30 – 15.00 Evaluating written assessments of practical work – a taxonomy
F. Wilson[1], N. Wade[1], S. Shaw[2], S. Hughes[2], S. Mattey[1]
[1]OCR, United Kingdom
[2]Cambridge Assessment, United Kingdom

Practical science work is a core component of science education, and is used not only to support the development of conceptual knowledge, but to enable students to develop a wide range of skills. As a result, many different models are used to assess practical science, ranging from coursework projects to written questions in examinations. For any mode of assessment, it is important to establish a clear understanding of what skills and knowledge are assessed. Furthermore, particularly when assessing a complex domain such as practical science, it is necessary to establish a clear and detailed framework to support the setting and evaluation of assessments, and which can be used to compare assessments across different educational stages and used in different contexts. Currently there is no established method for evaluating practical science assessments in this way.

In this paper we focus on one indirect method for the assessment of practical science: written examination questions. We present a taxonomy for classifying written questions about practical work, and demonstrate its use in the evaluation of science examination papers used in England and internationally. We conclude with a discussion about further applications of the taxonomy.

15.00 – 15.30 Measuring digital literacy and how to set a performance standard for pupils in 4th grade

O.E. Hatlevik[1], G. Egeberg[2]
[1]Oslo and Akerhus University College of Applied Science, Norway
[2]The Norwegian Centre for ICT in Education, Norway

Information and Communications Technology (ICT) is an important topic in the globalised world. It is important for both teachers and pupils to develop digital literacy in order to participate in our society and to be conscious about how to identify and distinguish between documented and faked information. In the area of alternative facts, we need to teach our pupils how to find, use and spread trustworthy information. A framework of digital literacy was used as starting point for developing a test for digital literacy in 4th grade (age 9 – 10 years old). A sample of 36,000 pupils conduced the test, and a cut-off score was set in order to identify pupils having critical low levels of digital literacy. Preliminary findings indicate a unidimensional construct of digital literacy that is rather fair for both girls and boys across the country. However, we found differences between schools and differences in digital literacy across age. One main question is therefore to what extent teachers can use the results from the test when planning their own teaching.

14.00 – 15.30 Session F: Improvements to Test Development Processes
Vltava and Vistula, Tom Bramley

14.00 – 14.30 Optimization of the Assessment of Dutch as a Second Language

T. Lampe[1], A. Verschoor[1]
[1]Cito, Netherlands

The movement and integration of people across countries is the reason why the Dutch government constructed a national examination system to assess the mastery level of Dutch as a second language in the Netherlands. The system is based on the lower levels of the Common European Framework of Reference for Languages. To allow every candidate to show what language level of Dutch they master, multiple choice tests are constructed to assess reading and listening levels. A written test assesses the writing skills of the candidates. These tests are mandatory to gain access to the Dutch educational system.
The recent large numbers of refugees require multiple equivalent versions of the same test. To give each candidate an equal and fair opportunity to show his or her ability, these test versions must meet different specifications, based on both content and psychometric properties as well. Computer software for optimal test design was applied to construct a number of equivalent test versions. Test specifications, item properties, and the predicted psychometric properties of the test versions will be presented. Finally, empirical results are compared to the forecasted outcomes, to evaluate this construction procedure.

14.30 – 15.00 Linking and standard setting in examinations: a framework for distinguishing different approaches and relating different sources of information

A. Beguin[1], M. Van Onna[1]
[1]Cito, Netherlands

In examination and test programs, multiple versions of the examination/test are frequently constructed; for example, across different years. To enable meaningful comparison of results between versions, the standard must be the same across versions. Different educational

systems have varying approaches to setting these standards. The approach adopted depends on culture and historical choices made in the design and development of the educational system. Often, these choices were not made explicitly, but developed over time, as adaptions of existing practice. In this paper, we propose a framework for interrogating those aspects of an assessment which impact on the comparability of standards. This framework provides a basis for systematic comparison between the ways in which different examination systems have implemented approaches to the setting and maintenance of standards.

Our proposed framework addresses, for each standard setting procedure, the approaches which can be taken in five categories:

- test design;
- expert judgement;
- population performance in the examination/test;
- comparisons between populations based on the assumption of random equivalence;
- linking procedures using item-level data.

For each of these approaches, further elaborations are given and we consider examples of approaches which combine a range of sources of information.

15.00 – 15.30 Paper Layout and Candidates' Views on Typeset Clarity
G.J. Zahra[1], D. Pirotta[1]
[1]MATSEC Support Unit, University of Malta, Malta

The MATSEC Support Unit has updated its guidelines to paper setters to improve and maintain the formatting standard of its examination papers. This would allow candidates to get familiar with a common paper layout and reduce non-subject related knowledge. The question of which typeset to use for an examination paper, however, is compounded with several arguments and beliefs. Suggestions in reports and organisations' websites include Arial, Comic Sans, Georgia, and Times New Roman. Additionally, for candidates with dyslexia, organisations have suggested fonts like Helvetica, Courier, Arial, Verdana, Computer Modern Unicode, Comic Sans, Century Gothic, Trebuchet, and Calibri.
The MATSEC Support Unit has conducted two quantitative research projects aiming to shed more light on the issue of typeset as experienced by Maltese candidates. The following research questions were tackled:

- From a selection, is there any typeset that is preferred by candidates?
- Do candidates with dyslexia show different preferences?
- Do other factors, such as gender and/or age, influence one's preference to a particular typeset?

A total of 772 participants took part in the research project, which was itself divided into two parts. The second part of the research project accounted for differences in font size and used printed questionnaires.

14.00 – 15.30 Session G: Validity Studies for Educational Tests
Suite 1, Isabel Nisbet

14.00 – 14.30 Evaluating the construct validity of educational assessment designs in the context of UK high-stakes qualifications
Y. Bimpeh[1]
[1]AQA, United Kingdom

In many public examinations in the UK, assessments are developed according to a table of specifications, in which items are dictated by assessment objectives. The assessment objectives are set out in a regulatory framework by the government's Department for Education. They are hypothetical constructs that are not observed or measured directly. Often, assessment experts assume the validity of the assessment objectives rather than

establishing their validity in a formal way. This study investigates the construct validity of the assessment objectives in a test that uses structural equation modelling.
The structural equation modelling provides a framework to analyse assessment objectives that are measured through multiple items. It is an integrated statistical procedure that tests the measurement model and all related hypotheses at the same time. With this model, we can answer questions about the validity of assessment objectives, construct reliability and how different constructs are related.
We discuss theoretical principles, practical issues, and pragmatic decisions to help evaluate the construct validity of high-stakes assessments in the UK. We illustrate the method with application to the new AS-level Chemistry, Physics and Biology examinations.

14.30 – 15.00 Validation of the Selection Process (MMI and Questionnaires) Used for Medical School Admissions
A. Moshinsky[1], D. Ziegler[1], N. Gafni[1]
[1]National Institute for Testing & Evaluation, Israel

In recent years, the use of assessment centers to screen applicants to prestigious study programs has become more common. Assessment centers examine non-cognitive variables related to personal and behavioral attributes.

Since 2004, four Israeli medical schools have adopted an assessment center as part of their selection process. The center consists of eight short structured interviews (MMI stations) and a biographical questionnaire.

The main criticism voiced against this system, which is very expensive compared to other measurement tools – has been the absence of a controlled validation study. However, validating this type of screening process is complicated because of issues involved in defining the qualities of a good physician and devising ways to measure those qualities. Moreover, hospitals are reluctant to allow researchers to collect statistical data on physicians who have completed their studies and moved on to residencies or work in the wards.

Despite these challenges, a validation study investigated the relationship between applicants' assessment center scores and their performance on OSCE (objective standardized clinical examination) stations as sixth-year medical students.

As expected, there is a significant correlation between assessment center scores and performance on OSCE stations. The study's structure and findings will be presented in full during the lecture.

15.30 – 16.00 Coffee

16.00 – 17.30 Session H: Maintaining Fair and Trusted Assessment in a Globalized World
Douro and Oder, David Haggie

16.00 – 16.30 Assessment in a “post-truth” world
M. Richardson[1]
[1]UCL Institute of Education, United Kingdom

This paper considers how assessment is perceived within a global “post-truth” culture, where “objective facts are less influential in shaping opinion than appeals to emotion and personal belief”. Global information sharing may be faster than ever before, but within a “post-truth” narrative, just how does the education sector maintain trust in educational assessments?

To expect people to believe in education, i.e. to consider it as having some kind of true value, requires the need to establish trust in how education systems are conceptualised, developed, presented and evaluated. However, in a "post-truth" world, the assessment community faces is challenged by systematic undermining of trust in public institutions and education is one domain where levels of mistrust are often high, and keenly felt by teachers, schools, pupils and assessment professionals alike.

High stakes examinations (both national and international) are publically criticised and perceptions of assessment are often characterised in negative ways. News and social media are dominated by "post-truth" discourses of suspicion unwittingly creating a "post-trust" attitude to assessment. This paper identifies key challenges faced by the educational assessment community in "post-truth" contexts and concludes with a discussion of potential strategies for challenging untruths, and establishing trust in assessment.

16.30 – 17.00 Fairness, justice and the role of assessment in a globalised world
I. Nisbet[1]
[1]*University of Cambridge, United Kingdom*

This paper will analyse the relationship between fairness of assessment and fairness (or justice) in society – in national and international contexts. What is meant by each and how do they relate? What is a fair international assessment? Can an assessment be fair but its outcomes be unfair because of wider societal factors? Or can an assessment be unfair if its results are modified to achieve a socially just outcome? Is there a meaningful concept of international social justice which is relevant to assessment? The paper will contrast two paradigms in thinking about assessment: as a contest and as a judgement of proficiency, with a danger of assuming the "contest" paradigm for all assessments.
The paper will then apply these concepts to issues of international portability of assessment outcomes, taking as examples the use of assessment results for university/college selection and decisions by national professional regulators on whether holder of qualifications gained overseas are fit to practise in the regulated country. To whom is fairness owed in these contexts? Is it fair to give preferences to learners in the home nation/state in allocating rationed educational opportunities?

17.00 – 17.30 High-stakes assessments in the transition from primary to secondary education in Northern Ireland: school level policies and children's views and experiences.
L. Henderson[1]
[1]*Queen's University Belfast, United Kingdom*

Admission to academically selective grammar schools in Northern Ireland is mediated by two different, privately operated, high stakes assessments. These assessments, due to a chaotic policy environment, are not subject to the same regulation as other high stakes assessments. No information about test outcomes, or their use in grammar school admissions, is made available by the testing organisations.
This research considers the particular social and ethical implications of the use of unregulated high stakes assessments at the transition to secondary level education. The chosen methods allow insights into the policy and practice of school transfer arrangements, as well as the views and experiences of children. Firstly, a documentary analysis of school level policies demonstrates differences in how test outcomes are used in informing school admissions decisions. Secondly, a survey of 1300 transition age children, designed in collaboration with child research advisors using a children's rights based approach, shows differences in children's experiences of school admissions procedures and admissions decisions.
The current assessment arrangements have become an accepted part of transfer to secondary education. However, there is evidence that the performance of school choice, engagement with the transfer tests, and school admissions decisions are mediated by socio-economic status.

16.00 – 17.30 Session I: Evaluating Teacher Assessment Practices
Danube, Andrew Boyle

16.00 – 16.30 Comparing teacher assessment practices of an engineering design challenge across countries

E. Hartell[1], G. Strimel[2], S. Bartholomew[2]
[1]KTH Royal Institute of Technology, Sweden
[2]Purdue University, United States

This paper reports the commonalities, and differences, in teachers' assessment practices in middle-school technology/engineering (TE) education, in Sweden, the United Kingdom, and the United States of America – three countries where STEM-education is strongly emphasized among stakeholders while also challenged in terms of teaching and learning opportunities for young pupils in school.
Even though TE education differs to some extent between different countries, open-ended design challenges are very common in STEM education programs all over the world. However, due to the open-endedness of design problems they are challenging to assess with reliability. In a globalized world this can be especially true when assessing work internationally as student expectations and teacher values may differ from country to country. Adaptive comparative judgment (ACJ) has been proven to provide valid, reliable, and feasible results for the assessment of open-ended design problems in TE education in several countries. However, the potential for ACJ, as a tool for international collaboration in assessment, has not yet been addressed.
This international comparative study utilized ACJ, as a tool, to investigate teacher assessment practices of student design challenges across countries.

16.30 – 17.00 The use of external and teacher assessment in school self-evaluation in Georgian schools: how organizational culture and trust towards teachers affects schools' assessment policy

N. Andguladze[1, 2], T. Bregvadze[1, 2]
[1]National Assessment and Examination Center, Georgia
[2]Ilia State University, Georgia

In some education systems, schools have been using student performance data for school self-evaluation. Schools normally use the data generated through external assessments or examinations. In Georgia however school self-evaluation and particularly the use of school-wide student performance data is a relatively recent phenomenon: school self-evaluation is mandatory for all schools, but regulations are rather loose; unlike many systems, national or local educational authorities in Georgia do not use student performance data to judge about school performance or rank schools. Nevertheless, the prevailing majority of Georgian school principals reported having used student performance assessment results for school self-evaluation. Some of the schools prefer commissioning student performance assessment to an external assessment provider, while others rely on teacher assessments. There is very little understanding of why Georgian schools choose one over the other. The present study uses mixed methods approach to examine the institutional characteristics associated with the school assessment policy, namely the choice between teacher assessment and external assessment in self-evaluation, attitudes towards the use of these two alternatives, and the differences in self-evaluation objectives, processes, and perceived outcomes.

17.00 – 17.30 How does technology assist (or not assist) teachers in their formative assessment practice?

A. Tolo[1], J. Chan[2], G. Stobart[3], A.T.N. Hopfenbeck[1, 2]
[1]University of Bergen, Norway
[2]University of Oxford, United Kingdom
[3]UCL Institute of Education, United Kingdom

One of the central features of the globalised world is the increasing reliance within education on the many forms of information technology available to teachers. This presentation will focus on how technology is contributing to formative assessment. It directly addresses the AEA sub-themes of Technological innovations in assessment and on New assessment formats.
This study reports an extensive literature review to outline what is known about the role technology is playing in assisting teachers' assessment practice and how this is being enacted in primary and secondary school settings. This review was conducted rigorously in several stages. The first was a literature search using a few keywords to lay a solid foundation for our understanding of the dominant research and trends in the field. The second stage of search yielded 1,584 articles, which were then investigated for relevance, producing 145 articles which were further coded. The analysis of the articles judged relevant is currently underway and the review outcome will be completed in summer 2017, making it timely for disseminations in the AEA conference in November 2017.

16.00 – 17.30 Session J: Establishing and Maintaining Standards 2
Amstel and Volga, Christina Wikstrom

16.00 – 16.30 Rank-order approaches to assessing the quality of extended responses

S. Holmes[1], C. Morin[1], B. Black[1]
[1]Ofqual, United Kingdom

The marking of external assessments must be reliable and the resulting rank order of candidates must fairly reflect their performance. We investigated alternative approaches to traditional marking of history essays to determine whether rank ordering can be applied consistently, and if so, whether it is more efficient to derive rank orders by simply placing responses in order or by using paired comparative judgement.
Examiners carried out both rank ordering and paired comparative judgement on the relative quality of responses, using a holistic construct derived by the most senior examiners to capture the qualities needed to answer the questions. Some anchor scripts were also selected which captured the construct in a relatively unambiguous way. All participants were trained to rank order using the construct, and then 60 candidate responses were ordered in an anchored rank ordering exercise. No ties were allowed. The same responses were also judged in an online paired comparative judgement task.
We compare mean rank orders from traditional marking, comparative judgement and anchored rank ordering tasks, as well as the inter-person reliability of the ranks for traditional marking and anchored rank ordering. Finally we compare the time taken to carry out the comparative judgement and anchored tasks.

16.30 – 17.00 The impact of government educational reforms on the maintenance of AS standards

K. Melrose[1]
[1]Ofqual, United Kingdom

In England, Advanced Subsidiary and Advanced Level qualifications ('AS' and 'A' Levels) are undergoing reform. These standalone qualifications were coupled such that an AS contributed 50% to the A Level but, from 2015, these qualifications are being decoupled. This has prompted concerns that students will not be motivated to perform as well as previously and that the uptake

of AS will reduce considerably with possibly only certain types of students continuing to take these qualifications. Such changes in the size and characteristics of the reformed AS cohort compared to previous cohorts threaten the maintenance of standards from pre to post reform due to the way standards are set in England. This research aimed to measure the extent of this threat by interviewing members of senior management in 17 schools in England. These schools were reducing their entry to AS over time, employing a range of entry strategies and influencing student subject choice. Schools were generally offering the same subjects currently but may reduce their offer if low numbers make running some subjects unviable. Some schools had seen a reduction in student motivation towards the new AS qualifications. The implications of these findings on the maintenance of standards will be discussed.

17.00 – 17.30 What causes variability in school-level GCSE results year-on-year?
S. Rhead[1], D. Patchett[1]
[1]Ofqual, United Kingdom

In 2015, Ofqual published a report on the year-on-year variability in General Certificate of Secondary Education (GCSE) results in England for schools, where variability was defined as the difference in the proportion of students achieving at least a grade C in successive years. Whilst most schools displayed little year-on-year variation, some schools displayed large year-on-year variation and there are some commentators that have expressed concerns that this is evidence that the comparable outcomes approach to standard maintaining might have a differential effect on schools operating in a more challenging context (such as a significant proportion of pupils from low socio-economic backgrounds). The aim of this paper is to use results from a number of GCSE subjects for the summer of 2012 to 2015 to explore empirically the relationship between centre variability and centre factors such as school type, change in entry; and student factors such as socio-economic status and first language. We will develop a multi-level logistic regression model to predict year-on-year centre variability.

16.00 – 17.30 Session K: Factors affecting assessment performance
Tiber, Yoav Cohen

16.00 – 16.30 Changes in self-reported test-taking motivation in relation to changes in PISA mathematics performance. Findings from PISA 2012 and PISA 2015 in Sweden
H. Eklöf[1], D. Reis Costa[1, 2], E. Knekta[1, 3]
[1]Umeå University, Sweden
[2]INEP, Brazil
[3]Florida International University, United States

In PISA 2015, after a decade of performance decline, Swedish students showed an increased performance in all literacy areas in PISA. Also, Swedish students reported a significantly higher level of motivation to spend effort on the test. The purpose of the present study was to describe the changes in test-taking motivation and mathematics performance between 2012 and 2015 in a Swedish context, and to investigate whether and to what extent the change in performance could be attributed to the change in test-taking motivation. When modeled in a multilevel SEM framework, findings suggested a significant effect of test-taking motivation on the change in performance and a non-significant effect of year on performance when modeled together. Findings thus indicate that increased test-taking motivation may be an important variable to consider when explaining the increase in performance, but also acknowledge that there may be other variables that may also be relevant to consider. Although the example in the present study is local, the issues raised have bearing also for the global research community, as the study aims to explore questions related to the true value of studies like PISA, how students may perceive these tests and how findings may be interpreted.

16.30 – 17.00 Factors related to reading trajectories of primary school children: results of a 3-year longitudinal study in Russia

I. Antipkina[1], E. Kardanova[2]
[1]Higher School of Economics (Russia), Russia
[2]National Research University Higher School of Economics, Russia

Reading is an achievement-defining skill but despite a number of theoretical works on the development of reading skills, Russian pedagogic science had lacked large-scale empirical studies on reading development. This study based on the data from iPIPS project is a first longitudinal study of Russian primary-school children designed in order to learn more about individual reading trajectories of children and assess the impact of individual, school and family factors. Reading skills, vocabulary, phonology and non-cognitive skills of 2195 children were assessed 3 times in 2014, 2015 and 2017: in the beginning of their schooling; in the end of their first school year; and in the beginning of the 3d grade. Students' parents filled in context questionnaires twice: first, regarding their education, income, family educational resources, child's pre-school experience, pre-school parental educational practices, and later, about parental involvement. Students' teachers also filled in questionnaires in the 1st and 3d grades regarding their experience, teaching materials and classroom practices. Last, in the 3d grade students themselves completed questionnaires regarding reading motivation and attitudes to school. This study looks at developmental trajectories of those children who came to school having no reading skills VS those who could read a little or read well.

17.00 – 17.30 Social-desirability in teachers' studies based on self-reports – the case of Russia

A. Kulikova[1]
[1]National Research University Higher School of Economics, Russia

Socially desirable responding (SDR) is the tendency of respondents to reply to a questionnaire in a manner that will be approved by other people. Teachers is a special group and their attitudes and opinions are socially sensitive. It can be expected that the answers of teachers can be strongly influenced by social expectations. The present study is aimed to discover what are the features of social desirability in teachers' responses in case of Russia.

The data from TALIS 2013 was used. The sample consists of 4000 teachers. The IRT theory was applied to build a SDR scale and check its psychometric properties. Then correlational analysis was conducted to estimate relationships between SDR and other characteristics.

The results show that SDR negatively relates to the difficulty of class contingent and positively to the age, self-efficacy and job satisfaction.
The results can be regarded as a signal for further analysis and interpretation of teachers' surveys based on self-reports. We should trust differently to responses from different teachers' groups in case of Russia. The most reliable groups are young teachers and teachers who work with difficult students.

16.00 – 17.30 Session L: Evaluating Test Quality and Features
Vltava and Vistula, Yasmine el Masri

16.00 – 16.30 Exploring test quality. An innovative approach to the display of item functioning using IRT

R. Coe[1], M. Walker[1]
[1]CEM, Durham University, United Kingdom

There is good evidence that the nature and content of national test items have a considerable influence on what is taught in classrooms and indeed upon how things are taught. It is important therefore, that items used in high stakes national tests are as good as they can possibly be and

that the information gathered via such test items is real information that reflects the measurement of underlying educational constructs.

The paper will present a method for exploring item functioning in tests that brings together complex information from item response theory, using the Rasch model, in a predominantly graphical manner. The method allows general users of test data to explore the functioning of individual test items at a level that might usually be accessible only to specialist test developers.

The paper presents a method of introducing technological and psychometric innovations in assessment and can point to the use of sophisticated validity evidence to support claims made about what high stakes tests can tell us.

16.30 – 17.00 Ensuring validity, fairness and equal opportunities in conducting summative assessment in Nazarbayev Intellectual Schools

S. Adikhanov[1], B. Yessingeldinov[1], A. Shilibekova[1], D. Ziyedenova[1]
[1]AEO NIS Branch Center for Pedagogical Measurements, Kazakhstan

Summative assessment procedure is one of the sources of systematic obtaining objective, transparent and comparable information based on the results of evaluation. Standardization of this procedure throughout Nazarbayev Intellectual schools is carried out by providing a range of conditions, such as, centralized development of specifications for summative assessment, internal and external expertise of summative assessment papers and functioning of a central archive of student papers.
The issue of developing test specifications for summative assessment is one of the conditions for maintaining quality and validity in assessment. Pedagogy is not static and the specifications for each test need to be continually reviewed and modified to reflect the current state of knowledge (A.S. Cohen & J. A. Wollack). This paper describes the stages summative assessment test specification development on the example of Nazarbayev Intellectual schools. In addition, comparative analysis of summative assessment results compiled on the basis of different specification formats, opportunities provided for the analysis of the results, teachers' opinion on the new format of test specification, as well as prospects for further development on this issue are presented in this paper.

17.00 – 17.30 Model parameters, interval scales, and the representational fallacy: Re-educating educational measurement

J. McGrane[1], A. Maul[2]
[1]University of Oxford, United Kingdom
[2]University of California, Santa Barbara, United States

A conceptual error lies at the heart of much of modern psychometrics: representations of reality (e.g., models, parameters, scales) are often conflated with reality itself (e.g., the psychological properties putatively measured by tests). Throughout the educational assessment literature, psychometric models and their mathematical properties are discussed as if they were either synonymous with or automatically applied to psychological attributes and their ontological properties. Entities such as model parameter estimates (e.g. estimates of "person abilities" in IRT models) are often treated and discussed as if they were entities with independent existence (e.g., students' actual abilities), or automatically refer to such entities. Further, the properties of these statistical entities (e.g., that estimates of "person abilities" can provide a linear ordering of persons, or, more strongly, can be arranged on a scale with interval properties) are commonly interpreted as if they automatically apply to the properties of the attribute as well, and thus debates around the appropriate representation of cognitive attributes are framed in terms of the assumption that the scale has interval properties, rather than the (actual ontological) assumption that the attributes exist and are

quantitative in nature. To resolve this fallacy, psychometricians must give primacy to educational theory in their models.

16.00 – 17.30 Session M: Studies in Admissions to Higher Education
Suite 1, Ayesha Ahmed

16.00 – 16.30 Adding a Writing Task to a University Admissions Test – An Evaluation of Short-Term Consequences
T. Kennet-Cohen[1], Y. Sa'ar[1]
[1]National Institute for Testing and Evaluation, Israel

The Psychometric Entrance Test (PET) is a standardized test used for admission to higher education in Israel. Until 2012, the PET consisted entirely of multiple-choice test items. In 2012 a writing task was added to the test.
At present, several points can be made regarding three core qualities of the test:
Reliability
The inter-rater (0.75) and test-retest (0.53) reliabilities of the writing task were as expected from a single task lasting 30-35-minutes. The reliability of the PET total score was minimally affected (test-retest reliability of 0.90).
Validity
Evidence of convergent validity (the correlation between the writing task and the average high school matriculation score was 0.45), as well as evidence based on the internal structure of the revised PET will be reported in the presentation.
Fairness
Women performed better than men on the writing task across all PET languages and especially in Arabic (Cohen's d for the gender difference was 0.35), even though women continued to perform more poorly than men on the multiple-choice sections. With respect to SES status, the gap between low-SES and high-SES examinees was smaller on the writing task (Cohen's d =-0.65) than on the multiple-choice sections (Cohen's d =-0.81).

16.30 – 17.00 Providing Validity Evidence for the Engineering Students Professional Competences Test (evidence from Russia and China)
E. Kardanova[1], D. Federiakin[1], P. Loyalka[2]
[1]National Research University Higher School of Economics, Russia
[2]Stanford University, United States

The aim of this study is to provide evidence regarding reliability, validity and cross-national comparability of assessment instruments that have been used for the Study of Undergraduate PERformance (SUPER) project. The main goal of the project is to investigate the quality of engineering education across multiple countries. More specifically, the goal is to assess and compare university student (levels and gains) within and across countries and examine which factors help students develop skills. In this study we seek to support validity and cross-national comparability of the professional competences test for electrical engineering students. A total of 841 Russian and 1,203 Chinese college students took the test.
We paid particular attention to differential item functioning (DIF) to provide evidence concerning the cross-national comparability of the test results and to ascertain the possibility of creating a common scale across the two countries. Also, we included response time in IRT-scaling. Results suggest that this allows to make ability estimation more precise than classical score-only estimation.

17.00 – 17.30 Towards routine use of validation studies to inform admission practices
P. Martinkova[1], I. Bartáková[2], A. Drabinová[1, 3]
[1]*Institute of Computer Science, Czech Academy of Sciences, Czech Republic*
[2]*Faculty of Education, Charles University, Czech Republic*
[3]*Charles University, Czech Republic*

Adequate selection of students to higher education is a crucial point for both the applicant and the institution. College admissions are usually based on written exams, interviews, and other criteria such as pre-admission grades, high-school leaving examinations, essays or recommendation letters. However, applied criteria may differ in their ability to predict success for different student populations. It is thus of high importance to routinely check for proofs of reliability and validity of the admission process.
In this work we describe the process of step by step improvement of admission process on the example of colleges and universities in Czech Republic. We first map the usage of various admission criteria and compare them to criteria used in similar programs abroad. Inspired by foreign studies we then study the degree to which the institutions validate their admissions. Finally, we discuss how boosting the routine performance of validation studies may be done by offering freely available software which includes training data and examples of analyses. We conclude by providing policy recommendations to improve the selection process to higher education in the Czech Republic and we discuss the possible generalizations of our results worldwide.

18.30 – 20.30 **Meet and greet for doctoral students** *(Skautský institut / Prague Creative Center Staroměstské nám. 4/1 (Old Town Square). Prague 1.)*

19.00 – 20.30 **Event for members holding accreditation** *(Skautský institut / Prague Creative Center Staroměstské nám. 4/1 (Old Town Square). Prague 1.)*

Friday, 10th November

Open Paper Sessions

9.00 – 10.30 **Session N: Marker and Marking Characteristics**
Douro and Oder, Rose Clesham

9.00 – 9.30 Putting a G-Theory approach to marking reliability through its paces
B. Smith[1]
[1]*AQA, United Kingdom*

During trials it became clear that the reliability values of items monitored in different ways were very different. Items monitored via seeding see large numbers of markers mark a handful of scripts (seeds), whilst in double-marking, two examiners mark many randomly sampled scripts. The presentation explains how a statistical manipulation rendered reliability statistics comparable across the two monitoring methods.

It also became clear that, in some cases, seeds were not representative of all responses (i.e. their mark distributions were very different). This raised concerns that the candidate variance estimated via G-theory methods was inaccurate. The presentation suggests substitution of the all responses' variance as a solution to this issue.
In addition, we discuss the interpretation of G-theory reliability statistics through real-life examples. Several caveats it is vital for users of the method to be aware of are covered, including small sample size and low candidate variance. The significant problem of unrepresentative data is discussed.

Finally, we cover how AQA, an English exam board, is putting these statistics to use via a continuous improvement process, and the training and resources that were needed to accompany this.

9.30 – 10.00 Giving G-theory marker statistics a context: comparison to classical measures and post-marking data

E. Harrison[1]
[1]AQA, United Kingdom

This presentation discusses the post-hoc application of G-theory to mark-remark monitoring data collected at AQA. The application is used to estimate the scale and impact of marker error on candidate marks and grades.

The G-theory approach is compared to a classical one, and illustrated using both 'poor' and 'successful' items and papers. The comparison indicates that G-theory can be used to produce equally understandable and useable statistics. The G-theory statistics also correlate well with the observed rate of mark changes from the post-marking enquiries about results service (where candidates can request that marking is reviewed), indicating that they give a reasonable prediction of the scale of marker error.

If the statistics were used operationally, test developers could use the information to spot good and bad practice in item design. They could then look to improve the design of items and mark schemes so as to reduce marker error in the future as part of a process of continuous improvement.

Limitations of the data collection are also discussed, together with possible future improvements.

10.00 – 10.30 Understanding the nature of marker disagreement, uncertainty and error in marking

C. Morin[1], S. Holmes[1], B. Black[1]
[1]Ofqual, United Kingdom

Recently, researchers have proposed a typology of marker disagreement and suggested four different categories of possible sources of disagreement. This paper presents the results of a study aimed at identifying the different characteristics that may lead to these four categories of marker disagreement. In order to achieve this objective, for seven papers from two subjects, groups of experienced examiners marked 50 clean scripts and reviewed 50 annotated scripts. The two sets of 50 scripts were constructed so that each script was marked by half the examiners and reviewed by the other half. For each examiner, the scripts within each set were presented in a random order and were marked using a bespoke online marking system. The marks awarded to each item were analysed and items were selected where different patterns of mark agreement/disagreement arose. During a series of one-day meeting, examiners were then asked to identify characteristics (e.g. in terms of the (i) response characteristics, (ii) item characteristics and (iii) mark scheme characteristics) where disagreement had been identified. The results will be discussed in light of the four categories of disagreement and possible ways of reducing marker disagreement will be suggested.

9.00 – 10.30 Session O: High Stakes Accountability Issues
Danube, Angela Verschoor

9.00 – 9.30 System monitoring and school accountability: Can both be adequately addressed by a single national assessment programme?
S. Johnson[1, 2]
[1]Assessment Europe, France
[2]Graduate School of Education, University of Bristol, United Kingdom

From a situation 30 years or so ago, when a mere handful of countries were benefitting from their own national assessment programmes, national assessment activity is today to be found in every part of the world, with rather few countries not yet involved. Its purposes are many and varied, as are in consequence its scale, form, manageability, cost and ultimate utility. While a principal purpose continues to be system evaluation through over-time monitoring, a relatively recent addition is school accountability. Some consequences of the growing school accountability agenda for national assessment design are evident in a number of trends. These include a move to the assessment of literacy (typically reading comprehension) and numeracy (cross-curricular skills), cohort rather than sample-based testing, the adoption of 'standardised test' approaches, and, where feasible, internet-based item delivery and automated marking. These moves raise new issues for national assessment, in terms of methodological appropriateness and the validity of the performance information ultimately provided. The presentation focuses on these issues, and in particular questions the degree to which the school accountability purpose can ever be satisfactorily met, and also the degree to which its introduction might be jeopardising the validity of over-time system monitoring in general.

9.30 – 10.00 Understanding the drivers behind early and multiple entry practices in Welsh high stakes examinations
R. Sperring[1], T. Anderson[1]
[1]Qualifications Wales, United Kingdom

In Wales, secondary school accountability is measured, in part, through results from externally assessed GCSE examinations. Some linear GCSE subjects offer multiple examination opportunities. Wales has seen a rise in the number of students being entered before the end of their final year in school with varying lengths of time spent studying the examination syllabus. Only the best grade achieved by each pupil contributes towards schools' accountability measures. This contrasts with countries such as England who replaced this approach with a policy that stipulates that only the grade from the first entry of a GCSE examination counts towards the calculation of a school's accountability measure. To gain a deeper understanding of the driving forces behind the rise in early and multiple entry for high stakes examinations in Welsh schools, Qualifications Wales conducted a piece of research to gather the views and experiences of those working in the education profession. Responses were analysed to identify emerging themes and motivations behind then culture of early and multiple entry. Key drivers included a feeling of pressure to improve attainment and consequently school performance measures, and the desire to maximise opportunities for pupils to succeed through the creation of "bespoke learning pathways" for individuals.

10.00 – 10.30 Approaches to predicting predictability of examination papers
S. Holmes[1], N. Zanini[1], B. Black[1]
[1]Ofqual, United Kingdom

Overly predictable papers can restrict the taught curriculum through teaching to the test and also threaten the validity of the assessment. This study took an innovative approach to identifying factors that influence predictability, by asking teachers to predict future

questions, and to justify their suggestions. Analysis of the reasoning behind their suggestions was used to develop a framework of predictability factors.
Experienced teachers predicted questions for the next paper from the specification they taught by considering past papers and the specification document. This was repeated for 6 specifications in total. They gave as much detail as they could on the reasons for their choices. At a subsequent meeting, some of the teachers for each specification then discussed the suggestions and produced a final set of questions and rationales.
The rationales and group discussions were analysed to produce the framework of predictability factors, which can be used to evaluate the predictability of an assessment, and also as guidance for assessment writers to reduce the predictability of assessments. In the summer the final set of predictions will also be compared to the actual summer 2017 examination papers to determine whether the predictions are accurate and the rationales valid.

9.00 – 10.30 Session P: Findings from Assessment Surveys
Loire and Elbe, Lesley Wiseman

9.00 – 9.30 Cambridge Progression – Teachers' Perspectives
M. Kuvalja[1], S. Hughes[1], S. Shaw[1]
[1]Cambridge International Examinations, Cambridge Assessment, United Kingdom

Cambridge International Examinations ('Cambridge') provides programmes of learning that progress from primary through to secondary and pre-university years. Cambridge offers a flexible approach in which schools can either offer every stage of the Cambridge curriculum, or they can focus on specific stages. Until now, however, it has been unclear as to how teachers perceive the progressive nature of the curriculum. This study aimed to explore teachers' perceptions of students' progression from one Cambridge stage of learning to the next.

A questionnaire was sent to 1500 Cambridge schools that offer at least two stages of the Cambridge curriculum. Teachers were asked to reflect on the extent to which their students came prepared for the current stage of their studies, and to report on students' readiness for starting the next stage of education. Teachers were also asked to identify the most important skills and knowledge that students need to become successful at each stage.

The majority of teachers reported that progression from one stage of the curriculum to the next reflects the nature of an aligned, instructional and progressive curriculum. Research findings are reported in relation to each educational stage and cultural context.

9.30 – 10.00 How consistent are male and female variances across national and international assessments?
L. Copping[1], S. Stothard[1], C. McKenna[1], G. Copestake[1]
[1]University of Durham, United Kingdom

Recent research on large, international educational data sets suggest that the "greater male variation hypothesis" is well supported. Males appear to be over-represented at the tails of ability distributions despite overall similarity in mean scores and the gradual closing of the attainment gap relative to females. This is particularly so in reading and mathematics and may help to explain why during the school years, males are often over-represented in high achieving groups as well as those with special educational needs. This in turn may explain why similar patterns are found across different social stratifications, with males being over-represented in high power professions and institutions. Here, we discuss our own programme of research using a combination of international, national and commercial data sets covering a range of different assessments. While it appears to be the case that males are more varied than females in literacy and numeracy assessments such as PISA, PIRLS and TIMMS, and across children's educational

careers, the pattern across other skills and subjects appears to be more complex and the effect sizes are far from homogenous. While the "greater male variability hypothesis" generally holds, there does appear to be some notable exceptions that warrant research attention.

10.00 – 10.30 School Readiness and Student Achievement: The role of formal and informal preschool education
T. Miminoshvili[1], N. Revishvili[1]
[1]National Assessment and Examinations Center (NAEC), Georgia

The paper describes the results of Trends in International Mathematics and Science Study (TIMSS 2015) as well as Progress in International Reading Literacy Study (PIRLS 2016) to present the following: (1) Comparative analysis of family environment factor (informal preschool education) and kindergarten enrollment effect (formal preschool education) for the formation of school readiness in preschool age has been conducted; (2) In this paper school readiness is discussed as a mediator variable between the inter-relation of student's academic achievement and formal and non-formal preschool education; (3) The aforementioned analysis is based on example of Georgia, along with other post-Soviet states and several East and Southeast Asian countries. Socio-economic as well as cultural context of these countries were taken into consideration when analyzing the existing differences.

9.00 – 10.30 Session Q: Political and Social Impacts of Assessment
Tiber, Rob Coe

9.00 – 9.30 Popular perceptions about the comparability of assessments in England. A tension between academia and the mainstream broadcast and print media?
G. Elliott[1], N. Rushton[1]
[1]Cambridge Assessment, United Kingdom

In England, academic research papers address the comparability of assessments over time. There are also television programmes and newspaper articles which address the same issue but take a different approach.
Situations where a middle-aged person takes a contemporary examination in order to comment upon its relative standard compared to examinations they took in their teens, or where a teenager sits an examination from the 1940s in order to contrast the experience with their own, can be seen as social experiments into comparability of examinations over time.
Such social experiments, conducted upon nationally important assessments, are relatively common in UK newspapers and are also to be found in broadcast programmes. Unlike academic studies, they reach a large target audience from many social demographic groups. Students, parents and teachers are key stakeholders in the educational assessment world and it is likely that their opinions on assessments are influenced by these media representations.
This presentation describes a selection of recently broadcasted and printed social experiments about the comparability of assessments in England. The advantages and the limitations of these experiments will be discussed in the light of how their derivation conforms to established research principles and how their presentation can influence key stakeholders.

9.30 – 10.00 Overcoming political and organisational barriers to international practitioner collaboration on national examination standard-setting

L. Gray[1, 2]
[1]*AQA, United Kingdom*
[2]*University of Oxford, United Kingdom*

This presentation will outline findings from an ESRC-funded project to investigate issues for researchers working in the politically sensitive area of national examinations. It will articulate the political and organisational barriers to such research and indicate ways in which individuals and organisations have overcome them to advance their national examination policies.

Examination boards find themselves in highly political environments in which their organisations and individuals who work for them can be scapegoated for political failings. Globalisation can drive standardisation, and often there are pressures to impose particular understandings of educational attainment, based on "evidence" about other systems' practices and their effects. This discourages transparency and reflection, which is detrimental to advancing understanding of theory, policy and technologies.

National examinations are inextricably linked to the wider educational culture in which they sit, so the issues are different in each setting. The presentation will suggest how exam board researchers can critically analyse personal and organisational practice, and the dominant policy and cultural environment within their own national setting. It will explore how exam board researchers can be more transparent about the challenges they face.

The presentation will be of interest to researchers, policy-makers and practitioners interested in facilitating transparency about assessment systems.

10.00 – 10.30 Developing a culture of research-informed practice by encouraging research uptake in an assessment organisation

S. Hughes[1]
[1]*Cambridge Assessment International Education, United Kingdom*

One purpose of AEA-Europe is the growth of members' knowledge and understanding of assessment. An international assessment organisation has put in place a number of mechanisms for improving research uptake to support the development of a culture in which assessment practice is research informed. This paper describes the mechanisms employed and the outcomes of the first phase of an evaluation of those mechanisms.
Mechanisms for improving research uptake include: dissemination of research outcomes in tailored formats; interaction between researchers and practitioners; facilitation through organisational support; and the use of rewards to incentivise research uptake.
A model for evaluating professional development was used to evaluate the impact of research uptake on assessment practice. This puts a focus on practitioners' reactions and learning; organisational support; practitioners' use of research knowledge; and the impact on students' assessment outcomes. Challenges identified relate to the capacity for managing research uptake, the targeting of practitioners with different roles and the need for a culture shift in which using research outcomes is a valued part of practitioners' roles.
Further phases of the evaluation will address impact at organisational level and build a longitudinal picture of the development of a community of research-informed practice.

9.00 – 10.30 Session R: Validation of Assessment Constructs
Vltava and Vistula, Paul Newton

9.00 – 9.30 Assessment tool validation research at Nazarbayev Intellectual Schools: student performance monitoring system in Mathematics
L. Issayeva[1], N. Dieteren[2], S. Crans[2]
[1]Nazarbayev Intellectual Schools, Kazakhstan
[2]Cito, Netherlands

Kazakhstan has started modernization of secondary education sector to educate functionally literate citizens. Nazarbayev Intellectual Schools (NIS) became an experimental ground for educational reforms including a combination of assessment methods supporting implementation of a new curriculum. Since the new curriculum was constructed following the spiral principle, it is crucial to track how well students seize what was previously taught to provide them with the relevant pedagogical support to improve learning. Student performance monitoring system in Mathematics has been being developed by NIS teachers and educational specialists together with the subject experts from Cito (the Netherlands) starting from 2011. The aim is to provide students and teachers with objective information about student progress in mastering NIS Mathematics curriculum. Objectivity of reported student performance data is guaranteed by standardized procedures for item construction, psychometric analysis, standard setting and describing ability levels. There is a substantial bank of monitoring items that are 'ready for test'. However, finalizing the development process of the monitoring system, NIS and Cito decided to conduct a quality assurance activity in order to critically validate the existing item bank. This study reports the results of item bank validation held by the experts from both sides.

9.30 – 10.00 Measuring Scientific Reasoning: Construct Validation of the Primary Scientific Reasoning Test (PSRT) using Rasch modelling
D. Ng[1]
[1]Oxford University Centre for Educational Assessment, United Kingdom

In the current globalised era, well developed scientific reasoning faculties are highly valued. However, it is unclear how scientific reasoning could best be assessed or if existing assessments are adequately measuring this proficiency. Progress in measuring scientific reasoning has been constrained by a limited understanding of the nature of scientific reasoning, which in turn impacts the development of suitable instruments.

The present study investigates the construct validity of a paper and pencil test – the Primary Scientific Reasoning Test (PSRT) when tested on 430 Singapore children who have completed their primary education (Year 6). The data was analysed using Rasch modelling. Findings revealed satisfactory psychometric properties on internal consistency, item quality, threshold-ordering and model fit. Overall, the results provided evidence that the PSRT was a valid assessment for evaluating the scientific reasoning abilities of primary pupils.

Finally, a case is made that two large-scale international assessments of science, which apply the Rasch measurement models to investigate items designed to measure complex scientific thinking, should publish analyses of how test-takers are reasoning with assessed content. Such information can help curriculum developers and teaching practitioners from different cultures better address common misconceptions, difficulties and the learning needs of their pupils.

10.00 – 10.30 Qualifications for the Construction and Built Environment Sector: a review of qualifications in Wales including comparisons with those in Germany, Canada, Australia and New Zealand
C. Taylor[1], G. Sparding[1]
[1]Qualifications Wales, United Kingdom

Following its ground-breaking review of the Health and Social Care sector in 2016, Qualifications Wales – the qualifications regulator for Wales – has undertaken a major review of qualifications, and the qualification system, in Construction and the Built Environment. Involving interviews with over 120 employers; focus discussion groups with over 800 learners; a technical review of the validity and reliability of specifications and learners' work and an online feedback mechanism, the review has also incorporated an in-depth comparison with equivalent qualifications in four nations: Germany, Australia, Canada and New Zealand. The review sheds light on some of the unintended consequences of historical features of qualification design as well as of some of the particular constraints that are imposed by public funding regimes. It questions whether efforts to provide clarity and consistency through the national and international conventions of qualification levels and standards may be providing unnecessary complexity and demotivating learners. The challenges of conducting effective work-based assessment in high-risk working environments are explored. The review identifies options for re-shaping the qualification pathways for learners entering this high-growth sector of employment.

9.00 – 10.30 Session S: Issues in Data Analysis
Suite 1, Naomi Gafni

9.00 – 9.30 Can teachers form an interpretive rating community? Quantitative and qualitative analyses
G.B.U. Skar[1], L. Jølle[1]
[1]NTNU, Norway

This paper reports on a reliability study of a national panel of raters, consisting of teachers from across Norway. Data consisted of ratings, interviews and live rating. The results were contradictory: the quantitative analysis revealed low reliability, but the qualitative analysis indicated shared perceptions on how to rate particular texts.

9.30 – 10.00 Pooling the totality of our data resources maintain standards in the face of changing cohorts
T. Benton[1]
[1]Cambridge Assessment, United Kingdom

Very often, we expect that the overall performance of candidates in one year should not differ too much from the performance in the next. Thus, given a large and stable cohort, major changes in outcomes may be indicative of a problem with standards. Going further, we may even use the assumption of stability to actually dictate where grade boundaries should be placed.

However, for international examinations we cannot assume stability. The candidates who choose to take a particular examination in Mathematics may come from different schools or countries from those entering next year. To address this, we consider how we might pool all of the data held by an assessment institution to create a single measure of candidates' ability that is calibrated across whichever assessments they have entered. By dealing with all assessments simultaneously we enlarge the size of the cohort considered and can now confidently use an assumption of stability to calibrate this measure across sessions. This provides a useful tool in the process of maintaining standards.

This paper will introduce new method to utilize data that has been pooled across many assessments and will demonstrate how it can increase the accuracy of standard maintaining in each individual examination.

10.00 – 10.30 Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests
T. Bramley[1], T. Benton[1]
[1]Cambridge Assessment, United Kingdom

Setting cut-scores representing comparable performance standards on different versions of the same test or exam is a problem faced by assessment agencies around the globe. If standard setting is conceived as a process whereby an abstraction (the performance standard) is made concrete as a cut-score on the raw score scale of a real test, then carrying out a standard-setting exercise on two tests is conceptually closely related to IRT true-score equating where score points on two tests corresponding to the same latent trait location are deemed equivalent. Noting that judge estimates of item difficulty typically correlate about the same with actual difficulty as empirical difficulty estimates based on very small samples (N=3), we compared, by simulation, a small-sample non-equivalent groups anchor test equating method with an Angoff-based method for determining equivalent cut-scores on two tests. At typical levels of correlation of judged and actual difficulty (r=0.6), small-sample equating with N=90 was more accurate than Angoff-based standard-setting. However, using a weaker anchor test or clustered sampling made the equating method similar to or worse than the Angoff-based method (depending on the cut-score location). We discuss implications for testing scenarios where these two approaches are likely to be feasible options.

10.30 – 11.00 Coffee

11.00 – 12.00 Discussion Group 1 *(Douro and Oder)*

11.00 – 12.00 Standard-setting/maintaining and public trust in national examinations around the world: the effects of structural and contextual issues
T. Isaacs[1], L. Gray[2]
[1]UCL Institute of Education, United Kingdom
[2]AQA, United Kingdom

Exam standards are part of a broader picture of educational and curriculum standards. This discussion group will explore that relationship. Through innovative poster presentations of six national standard-setting systems, it will exemplify how assessing the achievement of curriculum standards is powerfully enacted through structures and processes for standard-setting/maintaining in curriculum-related end-of-school examinations.

Educational cultures differ across jurisdictions, permeating assessment structures and processes in idiosyncratic ways. In standard setting, a key question is who has the power to set standards? Within any given national standard setting system, the number, nature and status of bodies involved and how they relate to each other, determine key features of that system. The way that the responsible bodies interact with wider stakeholders (such as examiners, teachers, parents and candidates), and how this changes over time, also has major effects on standard setting systems.

Through presentation of posters which outline standard-setting structures in six national systems, this discussion group will delineate key similarities and differences in cultural and contextual issues in the countries presented, and will provide a rich vehicle for exploration of the effects of these on standard-setting systems.

The discussion group will be of interest to researchers, policy-makers and practitioners interested in assessment standards.

11.00 – 12.00 Discussion Group 2 *(Danube)*

11.00 – 12.00 Using didactic tools for studying the content of mathematical assessments: an example with external and internal assessments in primary school classes in France

N. Grapin[1, 2], N. Sayac[2]
[1]LDAR, France
[2]University Paris-Est Créteil, France

Whatever assessment, before analyzing its results and interpreting, we strongly believe that studying the content of the test is primordial. In the framework of didactics of mathematics, a priori analyses are currently involved for studying mathematical tasks, relying answer process with teaching, but also with several characteristics of the task itself. We will present first a methodology, developed in a didactical framework and based on a priori analyses for analysing test content in a specific domain. Test tasks are successively studied one by one and after, over the whole test: we observe particularly the cover of the domain by the assessment tasks (is there any duplication? or any lack?) and the repartition of tasks according to their complexity. In a second part, we'll use this methodology for studying the content of two types of assessment: a national large scale assessment at the end of grade 5, and several internal assessments proposed by teachers for their students at primary school (grade 1 to 5). Lastly, with example of items extracted of TIMSS, we'll explain how such a methodology could be used to compare assessment test content and to interpret results in a comparative perspective between European countries.

11.00 – 12.00 Discussion Group 3 *(Amstel and Volga)*

11.00 – 12.00 National assessment design in this accountability era

S. Johnson[1, 2], J. Füeg[3], L. Munro[4], J. Strakova[5]
[1]Assessment Europe, France
[2]University of Bristol Graduate School of Education, United Kingdom
[3]EDK, University of Bern, Switzerland
[4]Scottish Qualifications Authority, United Kingdom
[5]Charles University, Czech Republic

National assessment activity is today to be found in every part of the world, in developing as well as developed countries. Its purposes are many and varied, as are in consequence its scale, form, manageability, cost and ultimate utility. The principal purpose continues to be system evaluation through outcomes monitoring: i.e. tracking the measured achievement of target student populations and subpopulations. Relevant learning conditions and environments within and outside the classroom are typically explored, to contextualise any observed change in the national picture of attainment. A relatively controversial, and more recent, purpose for national assessment, and one that holds irresistible and growing interest among politicians, is that of school accountability.

Consequences for national assessment design of the growing school accountability agenda are evident in a number of trends, which raise new issues in terms of methodological appropriateness and the validity of the performance information ultimately provided. The session will begin with a formal overview of current national assessment activity worldwide, supported by brief illustrative contributions from colleagues active in the field in different European countries. The second half of the session will be an open discussion, focused around some important questions for reflection.

11.00 – 12.00 Discussion Group 4 *(Tiber)*

11.00 – 12.00 Aiding cultural responsive assessment of migrant students in a globalising school

G.A. Nortvedt[1]
[1]University of Oslo, Norway

Teaching, learning and assessment take place in highly cultural settings, and both educational and cultural contexts might differ substantially between home and host countries for migrating students. International and national large scale-scale assessments demonstrate an "achievement gap" between migrant and majority students in most countries (e.g. OECD, 2015). This gap tells a story of fewer life chances, which has democratic implications. Across the world, culturally responsive assessment is viewed as a potential tool to address current assessment issues relating to the needs of migrant students. Culturally responsive assessment can be defined as classroom-based assessment that acknowledges and respects learners' cultural background and approaches to learning as they strive for academic success. Addressing the need to link assessment more closely to teaching and learning activities, assessment for learning might be the best tool to address cultural responsive assessment because the teacher is more cognisant of the needs of the individual student. The purpose of this discussion group is to explore what cultural responsive assessment for migrant students might look like and what kinds of competences the teacher and teacher educator might need. The focus will be on formative assessment.

11.00 – 12.00 Discussion Group 5 *(Vltava and Vistula)*

11.00 – 12.00 PDC Discussion Group on Standards: Quality of Assessment / Assessment of Quality

B. Hemker[1], R.V. Olsen[2], P. Newton[3], A. Boyle[4], E. Kardanova[5], E. Papanastasiou[6], S. Berger[7]
[1]Cito, Netherlands
[2]University of Oslo, Norway
[3]Ofqual, United Kingdom
[4]AlphaPlus Consultancy Ltd., United Kingdom
[5]National Research University Higher School of Economics, Russia
[6]University of Nicosia, Cyprus
[7]University of Zurich, Switzerland

The general goal of AEA-Europe is to act as a platform for discussion of developments in educational assessment in Europe. One of its tools to promote test quality is the European Framework for Standards for Educational Assessment available on the AEA-Europe website. The Framework is based on five principles: the focus on educational assessment; the fit for a European environment; an emphasis on ethics / fairness and the rights of the individual; addressing essential quality aspects (e.g. validity, practicality); and to support learning, decision making, test development, and programme review.

Five years after its presentation in 2012, the time seems right for an evaluation of the Framework and its use. The core of the discussion group is devoted to questions such as:

- Who is using the current AEA-Europe Framework? And how?
- How useful is the current AEA-Europe Framework? Are there contexts in which it is more or less useful?
- How could we improve familiarity and the usefulness of the AEA-Europe Framework?
- Do we need a new AEA-Europe quality evaluation tool?
- What should a new AEA-Europe quality evaluation tool look like to have added value?

The discussion group is wrapped up by formulating next steps for the professional development committee.

11.00 – 12.00 Discussion Group 6 *(Suite 1)*

11.00 – 12.00 E-Assessment Special Interest Group: 'Realising the potential of e-assessment'
M. Ware[1], L. Wiseman[2], N. Gafni[3], R. Hamer[4], M. Richardson[5], J. Moody[6]
[1]SQA, United Kingdom
[2]Assessment Consultant, United Kingdom
[3]National Institute for Testing & Evaluation, Israel
[4]International Baccalaureate, Netherlands
[5]UCL Institute of Education, United Kingdom
[6]Education Consultant, United Kingdom

Whilst its potential has been recognised for many years, the use of technology in educational assessment, or e-assessment, has not yet lived up to its early expectations. Use generally continues to be limited in scale and scope and often confined to digitising existing approaches to assessment rather than to helping us to adopt new approaches that reflect the opportunity for a more fundamental consideration of what we value and how we assess it that e-assessment offers.

To help support and promote adoption of e-assessment AEA-Europe is establishing an e-Assessment Special Interest Group (SIG). The SIG will be launched at the 2017 conference with this Discussion Group, which will include:

a) An outline of the purpose of the SIG and an introduction to members of the Core Executive Group (CEG)
b) Short presentations from members of the CEG providing an overview of the use of e-assessment in their countries/institutions
c) Stimulated by the presentations, a chaired discussion focussed on the unused potential for e-assessment, and barriers to development
d) Discussion of how the SIG could help to overcome those barriers, and
e) Types of activities to achieve this
f) Explanation of the process for joining and contributing to the SIG.

12.00 – 13.00 General Assembly
Suite 1

13.00 – 14.00 Lunch

Oral Presentations

14.00 – 15.30 Session T: E-Assessment
Douro and Oder, Amina Afif

14.00 – 14.30 Using Technology to formatively and summatively assess science, technology and mathematics inquiry based competencies
R. Clesham[1]
[1]Pearson UK, United Kingdom

This session describes the technology aided component of a FP7 EU funded project which researched the effective uptake of formative and summative assessment for inquiry-based, competence oriented STM education in primary and secondary education across seven partner countries in Europe

As part of the project, an on-line platform was developed to support inquiry based competencies in science, mathematics and technology education. The platform was also designed to scaffold inquiry based education (IBE), using formative assessment approaches. The affordances of the platform were designed to be as adaptable and flexible as possible in terms of providing an array of conversational, planning, data, video capturing and analytical mechanisms and tools to explore IBE. The platform provided students and teachers with an integrated environment to facilitate the creation of ideas, plan, execute and evaluate investigations, and also collaborate and share their work through peer and teacher assessment.
The online platform was trialled across European schools over a period of 18 months. The platform will be demonstrated in the session, alongside trialling feedback and outcomes. Finally, the implications and challenges concerning the design and development of on-line formative and summative environments in STM education for in-country policy makers and educational stakeholders will be discussed.

14.30 – 15.00 Accessibility for All Learners in a Computer Adaptive Test
S. Maughan[1]
[1]AlphaPlus, United Kingdom

The move from paper-based to computer-based tests is a topic of much debate: what advantages can the computer afford over paper-based equivalents, are the results comparable to the results from the paper-based format, can the technology be used to personalise the experience for the learners and so on. The Government in Wales is introducing a suite of computer adaptive tests for learners from age 7 to age 14 in reading and numeracy. Welsh Government is concerned to ensure that the tests are accessible to as many learners as possible. A programme of research has been conducted including: stakeholder interviews, reviews of academic literature, expert interviews and discussions with experts who have introduced similar assessments in other jurisdictions, into the most effective means of supporting the accessibility of computer-based tests. In this presentation, we will describe the research findings and the recommendations that were made for improving the accessibility of the computer adaptive tests in procedural numeracy. We will also show some examples of design and modifications that have been suggested and how they will be incorporated in the new tests.

15.00 – 15.30 Assessment in the era of Big Data: using data science to better identify students' strategies
P. Arzoumanian[1], T. Rocher[2]
[1]Ministry of Education, France
[2]DEPP, France

The culture of assessment is different from one country to another. French educational system is still influenced by classroom evaluation. Teachers are responsible of their ways to proceed to summative or formative assessments.

Standardized testing often suffered from its lack of feedback in the learning process. To improve this feedback, we need to be able to observe the actions realized by the students during the test.

Computer-based assessments give us the ability to go beyond the students' response. We can then collect the actions realized by the students to find the answer to the item.

We then analyze the collected data, to figure out the strategies developed by the student to solve the item. By characterizing the students' error, the assessment can allow the teachers to better understand the cause of the error, and then to adapt the learning process.

We realized this experiment on three interactive items of mathematics. They were included in a national assessment for students in 9th grade. We are presenting in the paper the results of this experiment.

14.00 – 15.30 Session U: Interpretation of International Survey Data
Danube, Theo Eggen

14.00 – 14.30 How to combine national and international assessments to diagnose the difficulties of the school system: the case of TIMSS grade 4 in France
M. Le Cam[1]
[1]*Ministry of Education, France*

In November 2016, the results of TIMSS 2015 (Trends in International Mathematics and Science Study) were released by the IEA. These international results have produced a shock in France, which ranked in the last place of the 22 countries or geographical entities of the European Union participating in the survey at the 4th grade. In parallel France has many national tools for measuring students' mathematical skills. The challenge is to combine national and international results to effectively diagnose the weaknesses of primary school mathematics education in France.
At the same period as the publication of the TIMSS results, a national consensus conference was held in France about how to learn mathematics in primary school. It ended with recommendations by a jury of practitioners after several experts' presentations. The results of the national evaluations have largely served as a basis for the debates and discussions at this conference. From the results of national assessments and of TIMSS 2015, it is possible to investigate learning issues in the field of numbers and calculus. In France, the results of standardized assessments, both national and international, are becoming increasingly important in the decision-making of educational policies.

14.30 – 15.00 Asia-Pacific and Scandinavian assessment cultures: more similar than different?
T. Burner[1]
[1]*University of Southeast Norway, Norway*

I take Vietnam and Norway as examples of typical Asia-Pacific and Scandinavian contexts. Typically, we think of these two contexts as diametrically different to each other when it comes to assessment cultures. However, my claim is that despite highly different history and context, the two regions represented by these two countries, have currently more in common when it comes to assessment cultures. Through colonization and political upheavals, Vietnam has been influenced by ancient Confucian, French, American and Russian education systems. The Confucian teaching philosophy has had great impact on the teaching and learning in Vietnam. Collective spirit is highly valued, in addition to respect for harmony and effort, respect for teachers, 'face saving' and love of learning, retaining a powerful influence on teaching and learning. In Norway, there is a more individualistic approach toward learning. Similar to Vietnam, students hesitate provide feedback. Teaching and learning are in both places oriented toward examinations and marks. Similar to the Asia-Pacific context, lifelong skills are promoted in Scandinavia. In both contexts, various assessment methods are used to tap students learning, and the most visible changes can be observed in elementary schools.

15.00 – 15.30 Item non-responses patterns across countries in TIMSS 2015
E. Papanastasiou[1]
[1]University of Nicosia, Cyprus

Construct irrelevant variance due to test-taking familiarity or test-taking practices can impose threats to the validity of data interpretations from comparative international studies. Therefore, it is important to identify variations that might exist in test-taking practices from country to country, as well as within countries. The purpose of this study is to identify the variations that might exist in relation to a single test-taking practice, that of omitting item responses, as well as the examination of possible predictors of this practice, by examining TIMSS2015 fourth-grade data. The results of the study have found that there are significant variations between two countries that also have very distinct testing cultures. Moreover, item characteristics are also related to the ways in which students respond to various test items, and more specifically, to non-responses.

14.00 – 15.30 Session V: Assessing Hard to Measure Constructs 2
Amstel and Volga, Guri A. Nortvedt

14.00 – 14.30 Assessing group dialogue: what is good participation in group work and how can we assess this?
A. Ahmed[1], R. Johnson[2]
[1]University of Cambridge, United Kingdom
[2]AQA, United Kingdom

We report results of a study investigating how features of group work can be assessed. The inclusion of collaborative tasks in PISA indicates their global prominence, so it is important to consider how to facilitate teaching and assessment of these skills.

We aim to identify features of dialogue that:
• are important for good participation in group work
• result in better outcomes of group processes
• can be assessed by teachers to inform teaching and provide useful feedback for learners

We filmed 15-year-old students participating in robotics tasks and collected teachers' observational notes and comparative judgements of performances in the talk, problem solving and social elements of the interactions. Our analysis involves dialogue coding and is guided by a socio-cultural perspective in which solutions arise through the co-construction of meaning.

Assessing collaboration is a challenge in the political climate in England: teacher assessment is not highly trusted and school accountability is based on external value-added measures. Tensions between teacher assessments and external exams are particularly apparent in the assessment of skills that are critical for group work. We hope a better understanding of how to assess group processes can lead to collaborative skills becoming more valued in our curriculum.

14.30 – 15.00 An Exploration of the Nature and Assessment of Student Reflection
S. Shaw[1], M. Kuvalja[2]
[1]Cambridge Assessment, United Kingdom
[2]Cambridge International Examinations, United Kingdom

Cambridge International Examinations (Cambridge) aims to develop not only subject-specific knowledge, but also encourages students to acquire vital skills important for further study, professional development and life in general. For example, Cambridge learners are

encouraged to be confident, responsible, reflective, innovative and engaged intellectually and socially. This aspiration is reflected in the syllabuses and assessments, offered by Cambridge, which assess, among other skills – students' reflection. The integrity of a Cambridge tests depend to a large extent upon a comprehensible understanding and articulation of the underlying abilities or construct(s) which they seek to embody. If these construct(s) are not well defined then it will be difficult to support the claims a test developer may wish to make about the usefulness of the test. It is, therefore, of crucial importance to understand the concept of reflection and to identify specific behaviours which represent the concept in order for reflection to be operationalised for assessment purposes. Cambridge has developed syllabuses which are specifically created to prepare students to think critically and to develop reflective thinking for students from 5 to 18 years of age. This presentation will give an overview of the literature on reflection as part of the student's learning process.

15.00 – 15.30 National assessment of citizenship in Flanders: knowledge and attitudes of students at the end of secondary education
L. Willem[1], M. Vandenbroeck[1], E. Ameel[1], D. Van Nijlen[1], R. Janssen[1]
[1]KU Leuven, Belgium

To connect to trends in international educational research, the Flemish government wants to asses more "non-traditional" topics. Therefore in March 2016 a national assessment in civic and citizenship education was organized. The outcomes of this national assessment are two-sided: knowledge and skills as well as attitudes were tested. The main objectives of the assessment were to map out differences between different groups of students in performance and attitudes with regard to citizenship, to investigate school differences in civic knowledge, skills and attitudes of their students, and trying to explain these differences by pupil and school characteristics. The results indicate that there are large differences between groups of students in both knowledge and attitudes. Students from the vocational track for example, have significantly lower scores on both the knowledge-tests and the attitude-scales. At the level of the school, we find very few differences between schools, especially when we take the tracks into account. Differences between students and schools are mainly explained by student-characteristics. We find very few effects of variables at the school level.

14.00 – 15.30 Session W: Features Which Impact on National Test Results
Loire and Elbe, Frans Kleintjes

14.00 – 14.30 Validity issues in educational assessment – should subscores in national tests be reported or not?
A. Lind Pantzare[1]
[1]Umeå University, Sweden

In the Swedish criterion-referenced school system teachers are trusted to teach, assess and grade their students. The grading is high stakes since the grades are used for admission to higher education. There are national tests for some of the courses. However, the national tests are not final examinations and the results from the national tests are not decisive in the grading. The main aim with the national tests is that they should support fairness and equality when assessing and grading the students.

In 2011, new syllabuses for upper secondary school were introduced. In mathematics the most obvious and visible change was an ambition to set an even larger focus on competencies instead of content.

The result on the Swedish national tests are reported in the form of a test grade. It has also been taken for granted that the only reasonable is to report the total result and nothing else. However, there has been an increased demand to not only report results based on the total

score but also report subscores connected to the competencies. The question is if the national tests are developed in that manner that it is possible and relevant to report subscores based on the competencies.

14.30 – 15.00 Analysing multidimensional ordinal data in attainment-referenced assessment

A. Scharaschkin[1, 2]
[1]AQA, United Kingdom
[2]University of Oxford, United Kingdom

Assessment culture in England has resisted the imposition of largely closed-form or standardised testing models, similar to the SAT in the US, with respect to national high-stakes summative assessments. The summative valuation of students' responses to tasks requiring constructed responses, such as essays, performances, artwork, etc., forms a substantial part of these assessment procedures. The assessments are curriculum-embedded, and it is necessary to demonstrate that students' overall results reflect the intended assessment objectives: that performances which are graded 'C', for instance, tend to exemplify the qualitative features that are supposed to be associated with 'a typical grade C performance'. In this regard, public examinations in the UK have been characterised as 'attainment-referenced'.
This presentation seeks to model features (construct-relevant attributes) of performances in attainment-referenced assessment as mappings that associate with each performance an ordinal 'value' (not necessarily a number, but a member of a partially-ordered set). It examines the prospect for using an analogue of principal component analysis for ordinal data to appraise the extent to which particular qualitative features are present in different classes of performances. Applications to the design of marking and grading procedures will be discussed.

15.00 – 15.30 The study of the gender factor influence on the results of students' learning achievements at Nazarbayev Intellectual schools

O. Mozhayeva[1], A. Shilibekova[2], Z. Rakhymbayeva[3], C. Jongkamp[4], F. Kamphuis[4], F. Kleintjes[4], A. Jandarova[2]
[1]NIS, Kazakhstan
[2]AEO Nazarbayev Intellectual Schools, Kazakhstan
[3]Nazarbayev Intellectual Schools, Kazakhstan
[4]Cito, Netherlands

The issue of gender identity in the education system in Kazakhstan is characterized by inconsistent and often inaccurate data from various sources. These data are difficult to translate into information that can be used in planning the educational process.
Gender influences student performance because of the difference in use of "self-regulatory behavior strategies" by boys and girls (Ablard, Lipschultz, 1998).
In the practice of students selection process in the 7th grade of Nazarbayev Intellectual Schools (NIS) no significant differences in the results of girls and boys are observed. This may be the result of preliminarily testing of selection test items for the identification of gender differences in the predictive validity of the results. However, the issue of the absence or presence of the gender factor influence on the results of students' educational achievements in the process of further training in NIS remains unexplored.
The aim of the study is to study the gender factor influence on students selection test results in the 7th grade and the trajectory of their further study at NIS. For the analysis, quantitative indicators were collected for students who were selected in 2013 and studied at Intellectual schools for 3 years (more than 6000 respondents).

14.00 – 15.30 Session X: Item Writing and What Affects It
Vltava and Vistula, Bas Hemker

14.00 – 14.30 A culture of question writing: How do question writers compose examination questions in an examination paper?
M. Johnson[1], N. Rushton[1]
[1]Cambridge Assessment, United Kingdom

At AEA-Europe 2016 we reported on a project that studied how question writers composed individual examination questions. We wanted to extend this question writing model to consider how question writers developed questions when writing a whole examination paper. To do this we observed the question writing practices of six question writers of examination papers that included a variety of response types. It was hypothesized that writers who develop questions for such papers may have a mental model that encompasses other items when they write individual questions.

Each participant wrote an examination paper whilst thinking aloud, and a researcher observed this task to capture the nature and sequence of elements of the activity. The writers were also then interviewed. For our data analysis we adopted a sociocultural approach that encouraged us to take into account both the cognitive and the social dimensions of professional question writing practice. A specific area of interest for analysis was the consideration of how social perspectives were evident within the writing model, and how these may have an important role in quality assurance. The project gives insights into the lived experience of question writing and the outcomes are helpful for the training of new writers.

14.30 – 15.00 Do translated items perform the same way? The experience of assessment in a bilingual country
M. Hogan[1]
[1]WJEC, United Kingdom

Background: Wales is a bilingual country with candidates sitting exams through either the English or Welsh languages. For most subjects an original paper is professionally translated so that there are two papers and candidates can pick which one to sit. Considerable effort is exhorted to ensure that the translation produces questions in both mediums which are of identical difficulty. However until very recently no work has been done to use the data produced during marking to empirically measure differential item functioning.

Methods: This work uses differential item functioning analysis for polytomous items to compare the performance of candidates on different items who took the papers through different languages. The analysis is performed across a number of different subjects.

Results: Some of the work to be presented will be based on summer 2017 examinations. However pilot work shows that variation between languages, whilst statistically significant, has tended to be well under 5% of the marks available for an item and implies that these effect sizes are not educationally significant. The results are in the context of an evolving wider DIF strategy whereby flagged items are submitted for qualitative review.

15.00 – 15.30 Alternative uses of examination data: the case of English language writing
L. Chambers[1], F. Constantinou[1], N. Zanini[1], N. Klir[1]
[1]Cambridge Assessment, United Kingdom

Examination boards are unique in that they have access to examples of student writing that span attainment levels and, if stored, can span time. This resource can be harnessed to generate valuable insights capable of informing education policy and practice. One use of this resource is

to investigate students' use of written language over time. This study compared two corpora consisting of 100-word extracts of students' examination writing from 2004 and 2014. The corpora consisted of 858 narrative texts written by 16-year-old students as part of a government-regulated high-stakes examination in English Language.
The aim of the research was to examine whether the formality in students' examination writing changed over time. The analysis focused on a number of linguistic features serving either as discriminators between spoken and written discourse (e.g. lexical variety, lexical sophistication), or as markers of informal electronic communication (e.g. abbreviations, omitted stops, non-capitalised sentences). The statistical analysis carried out between the two corpora indicated that, overall, the texts produced in 2014 were characterised by less formality than the texts produced in 2004, particularly for lower attaining students. These findings suggest that students should be supported in developing their awareness of context-appropriate written language.

14.00 – 15.30 Session Y: Innovations in Technology to Support Test Development
Suite 1, Martin Drnek

14.00 – 14.30 Improving Assessment Literacy among Teachers and the General Public Using MOOCs
A. Allalouf[1], N. Friedman[1]
[1]NITE, Israel

In many countries, the field of educational measurement and psychometrics is underdeveloped in academic frameworks. To redress this, a committee was set up, which recommended that NITE develop a certification program in psychometrics. The program consists of six courses in the areas of testing theory, statistics, and research methods; assessment development; and societal effects of testing. This presentation will focus on the massive open online course (MOOC), "Developing Measurement and Assessment Tools," which is designed to be relevant and attractive to teachers and relevant professionals and will be open for all. It comprises 60-videotaped modules as well as six frontal exercises, and is taught by 25 experts. The course is divided into seven sections: (1) introduction (reliability, validity, fairness); (2) developing open- and close-ended items; (3) developing assessments in educational contexts (national and international); (4) evaluating specific competencies; (5) assessment centers for non-cognitive traits; (6) developing questionnaires; and (7) other topics, including test translation, and technological and psychometric innovations. We believe that presenting the course will enhance and expedite international cooperation regarding the development of teaching tools, and increase awareness of these issues among teachers, policymakers, and the general public.

14.30 – 15.00 The role of technology in supporting innovation in assessment
D. Haggie[1]
[1]GradeMaker Ltd, United Kingdom

This paper presents the experience of three assessment bodies who are using new exam authoring technology to support innovation in their test development processes and services to centres. The first example is of an awarding body operating in a context in which security pressures are paramount, who is using technology to enable a shift from test to item development and extend the use of pre-testing. The second example is of an authoring team who, when adopting digital processes for test development, began to explore author specialisation, enabling them to focus authors in specific areas of the curriculum or item types to reflect their skills. The final example if of a board adopting authoring technology to support longer term plans to introduce eTesting. The exam authorities are presented without attribution. It is argued that while reviews of technology focus on the innovation directly offered by technology itself, much of the impact of technology is in the way it liberates exam teams to innovate in their own practice.

15.00 – 15.30 Flexible Examination and Equality: how to Turn two Foes Into Friends
A. Verschoor[1], T. Lampe[1], E. Roelofs[1]
[1]Cito, Netherlands

In our ever faster changing society, the call for flexible examination systems gets louder and louder. Re-use of items, multiple parallel versions, and test equating before the examination takes place are a few of the tools to make flexible examination possible. But, if done not carefully, they endanger the best practices related to the development of fair examinations.

In our paper, we propose several automated test assembly (ATA) models that speed up some phases of test development, while at the same time they make sure that equality and security can be guaranteed even better than traditional procedures. In three examples of examination systems we will show how the use of ATA provides us with examination variants of higher quality, better item use and related lower risk of items getting known to the candidates, and more possibilities for test equating than in the recent past: the Dutch digital Central Examinations, the theory examinations for acquiring driver's licenses, and Dutch as a Second Language (DSL). For the Central Examinations, up to 12 versions are constructed annually, while for DSL the number of versions will be expanded from 8 to 40. The Driving Licensing Authority has 10 parallel versions that are replaced weekly.

15.30 – 16.00 Coffee

16.00 – 17.30 Session Z: Impacts of National Testing
Douro and Oder, Sandra Johnson

16.00 – 16.30 The influence of the National Reading Tests on teaching and learning of reading strategies – a Welsh secondary school case study.
J. Nicholas[1]
[1]National Foundation for Educational Research, United Kingdom

In 2013, in the face of concern about low literacy standards, standardised reading tests for 7-14 year olds became statutory in Wales. International research literature shows polarised views about the effect of high stakes tests on education standards, while suggesting that a combination of teaching reading strategies, offering feedback on performance and the promotion of reading, increases pupils' self-efficacy and confidence levels. By means of a case study, this paper considers the influence of the tests on the teaching and learning of reading strategies and on pupils' attitudes towards reading. A sample of 302 pupils, aged between 12 and 14 years old, was surveyed, together with a sub-sample of focus groups and interviews with key staff. While pupils showed awareness of a variety of reading strategies, and reported high levels of confidence in preparation for the tests, there were mixed reactions towards specific whole school literacy sessions. Generally, older pupils' attitudes to reading were more negative. Furthermore, an emphasis on accountability had led the school to focus on test preparation at the expense of using the results diagnostically.

16.30 – 17.00 Are National tests mirroring school based assessment?
J. Radišić[1], A. Baucal[2], G. Čaprić[3]
[1]University of Oslo, Norway
[2]University of Belgrade, Serbia
[3]Institute for Education Quality, Serbia

Grounded on the idea of different assessment cultures across Europe and different functions external exam at the end of compulsory education and school based marks serve in the education system in Serbia we examine links between the exam score and school marks in Math with the aim to validate school marks based on the exam scores. In particular, to what extent criteria for

different school marks vary across schools? The analyses gather 12943 students from 104 schools using the 2016 exam data. Results show that while within a particular school students with lower marks typically have lower exam score, there are large differences across schools in their criteria for awarding certain school marks; indicating students with the same exam score might have different school marks. The findings suggest that the validity of school marks can be put in question. Furthermore, taking into consideration that school marks contribute with 70% in total enrollment score when students apply for enrollment in certain upper secondary program means that the current system favors students who attend schools with lower criteria in school assessment. Validity of school marks, and equity in the system are further discussed.

17.00 – 17.30 The introduction of national assessments in Norwegian higher education: Challenging views on assessment and autonomy in higher education
S. Hamberg[1], R.V. Olsen[2], K.C. Skåtun[1]
[1]Norwegian Agency for Quality Assurance in Education (NOKUT), Norway
[2]Center for Educational Measurement, University of Oslo, Norway

From 2014 and up to this date full scale trials of a few joint national exams have been conducted for the higher education programs in teacher education, nursing and accounting and auditing. Prior to this trial results from the exams in higher education suggested that institutions have their own idiosyncratic exam and grading practices. With the results of the nationally developed and graded exams it is now possible to study this bias explicitly. We provide results from a multilevel regression analysis modeling the degree to which local grades can be accounted for by the results in the joint exam, students and institutions GPA and other relevant characteristics of the students and the institutions. Furthermore, we present and discuss how the introduction of the national exams challenges long held views on assessment in higher education institutions, and related to this how values of autonomy in higher education are challenged

16.00 – 17.30 Session AA: Social and Cultural Issues in Assessment
Danube, Ronan Vourc'h

16.00 – 16.30 International and national periodic assessments: Thick as thieves
D. Van Nijlen[1], J. Denis[1], R. Janssen[1]
[1]KU Leuven, Belgium

One of the strongpoints of national assessments is the periodicity. Repeating assessments can inform policymakers not only on the overall performance evolution, but also on the evolution in achieving equity for different groups of students. In the present paper we present the results for the 2016 national assessment of mathematics at the end of primary education in Flanders, but we also show how interpretation of the results can be strengthened by framing them in an international context.
Mathematics had been assessed before in 2002 and 2009. Overall, between 2002 and 2009 performance was stable. However, when 2016 results are compared to 2009 for the majority of tests we see a steep decline in performance and the performance gap between boys and girls (in favor of the boys) significantly widened for four tests. These results seem to be in contrast with the overall good performance for mathematics of the Flemish students in TIMSS 2015. However, the evolution of the Flemish results in TIMSS tells us that Flanders is one of the few regions that does not show an increase in performance. Moreover, Flanders is one of the regions where the performance gap between boys and girls seems to widen.

16.30 – 17.00 Should there be a single assessment culture in a globalised world?
T. Oates[1]
[1]Cambridge Assessment, United Kingdom

This presentation will examine contrasting paradigms in assessment, and examine whether a single assessment culture is possible and desirable – the 'should' in the title. The analysis will draw on transnational comparison of a range of countries, looking not only at the forms of assessment (underlying models and practices) but also other factors (such as models of ability), which the analysis will argue are determining of 'assessment culture'. The analysis will include comparisons of the models and approaches in the large transnational surveys, and will argue that these are more contrasting then often is assumed. The analysis will pick up the issue of the construct focus of assessment, and will argue (on the basis of consequential impact) that variation in this is as important as the variation in measurement models and assessment paradigm, but should not be confused with them. It will be critical of convergence on a single model, but will argue that assessment development is not an eclectic free-for-all, but should be a process of principled application from a menu of assessment models.

17.00 – 17.30 East meets west: how social and cultural contexts can have different impact on high-stakes national assessments
A. Yessengaliyeva[1], N. Dieteren[2]
[1]PhD Sociology, Kazakhstan
[2]Cito, Netherlands

Exams, assessments, evaluations: testing was, is and will be the most widespread method of measuring knowledge and skills of individual learners in any society. But, culture of assessment can be diverse. In some countries, like Netherlands and United Kingdom, standardized testing has a long history and is really rooted. In most countries to the East of Europe this tradition is quite young.
In modern Kazakhstan the main high-stakes assessment is UNT (Unified National Testing). UNT was implemented thirteen years ago and until 2017 it combined final certification for secondary education and entrance examinations to higher education. The combining of two different purposes of testing has caused much discussion and loss of trust in the efficacy and reliability of the UNT. It also had different social and cultural consequences for the society, where a new generation of Kazakhstani with different views, ways of thinking and attitudes behavior developed. Since 2017 the UNT is split into a separate school-leaving exam and entrance exam for university.
Based on publications and own survey this paper will present an analysis of how social and cultural context can have different impact on validity of systems for central examination.
We will compare Kazakhstan, UK and the Netherlands.

16.00 – 17.30 Session BB: Evaluating Innovations in Assessment
Loire and Elba, Amina Afif

16.00 – 16.30 Introducing Progress Maps in the Czech Republic – Lessons Learnt
M. Drnek[1]
[1]Scio, Czech Republic

Scio as the largest local provider of assessment tools for schools has been long involved in researching and changing the culture of assessment in the Czech Republic. We observed that the form and method of assessment can have a great impact on both performance and outcomes, and that there is a great need for a better deployment of formative assessment practice. Learning progress maps are a tool for mapping students progress widely used e.g. in Australia, Ontario, etc. Scio, within the Learning Progress Maps project, took over the concept of the maps and enhanced it with standardized assessment tools.

One of the most exciting innovations in this Scio assessment are classroom assessment rubrics and projects. These rubrics are directly linked to the educational objectives formulated by maps, working both with feedback from the teacher and students self-assessment. As a whole, the system creates a space for providing individualized written and verbal formative feedback along standardized tools, and thus increasing the objectivity of assessment. In the conference paper, we will guide you through the features of the system and share our experience gathered over 6 years and from over 80 schools involved.

16.30 – 17.00 A review of the design and assessment model of a skills based qualification within the Welsh Baccalaureate.
K. Jones[1], T. Anderson[1]
[1]Qualifications Wales, United Kingdom

To prepare young people for life, further education, and the workplace, the teaching and learning of employability skills is a vital part of 14-19 education in Wales. This is facilitated through the Skills Challenge Certificate (SCC) qualification, which provides learners with an opportunity to develop and demonstrate skills such as Literacy, Numeracy, Digital Literacy, Planning and Organisation, Creativity and Innovation, Critical Thinking and Problem Solving, and Personal Effectiveness.

As part of an on-going programme of continuous improvement, the design and assessment model of this qualification is currently being reviewed. This review has collected evidence from a range of sources including focus groups with learners, semi-structured interviews with teachers and stakeholders, an evaluation of key materials produced to support the delivery of the SCC and a literature review of employers' skills needs. This evidence is being analysed within the principles of 'Constructive Alignment' to assess whether the qualification is 'fit for purpose'.

The findings of this review are still emerging but will be used in Wales to establish whether any developments should be made to the SCC. The presentation will reflect on how employability skills could most effectively be delivered and assessed, including consideration of the role of teacher assessment.

17.00 – 17.30 Exploring students' experiences of the Extended Project Qualification
C. Stephenson[1]
[1]AQA, United Kingdom

Current educational policy in England restricts the amount of school-based assessment within general qualifications, despite suggestions that some school-based assessments, such as those that are project-based, can enhance learning. Relative to traditional content-based teaching methods, project-based learning can boost students' academic performance by increasing engagement, self-direction and motivation and enhancing academic skills. Research into project-based learning and its effectiveness has examined its effects on academic performance within the same academic discipline as the project undertaken. However, recent evidence suggests that students undertaking a project-based qualification, the Extended Project Qualification (EPQ), also demonstrated enhanced performance in other subjects. This research contributes to the extant literature by exploring students' perceptions of the effects of undertaking the EPQ on their general academic performance. A qualitative investigation explored the experiences of 15 EPQ students using semi-structured interviews. Thematic analysis was used to analyse the data and identify emergent themes. This open paper reports on the resultant themes which illuminate the potential benefits of project-based learning. Importantly, the paper voices students' perceptions of the effects of school-based assessment in a nation where examinations are the preferred approach to educational assessment.

16.00 – 17.30 Session CC: Complex factors affecting measurement outcomes
Tiber, Lenka Fiťtová

16.00 – 16.30 Longitudinal analysis of the role of social context on motivation and perceived self-efficacy
L. Ben Ali[1], R. Vourc'h[2]
[1]DEPP B2 – Ministère de l'Education nationale de l'Enseignement supérieur et de la Recherche, France
[2]Ministry of Education, France

This communication enables us to study another aspect of the reproduction of social stratification as we can consider socioemotional behavior as partially mediated by the social context.
We focused on the evolution of students' motivation and perceived self-efficacy between the beginning and the end of lower secondary education. It is based on data collected from a panel of students who entered the 6th grade in 2007 in France. For this, we used data from standardized assessments held during the first and the final year of lower secondary education. In total, the population represents nearly 24,000 young people for whom we also have information regarding their social and cultural background. The aim is to verify with a longitudinal study combining cognitive results, social context measures and socioemotional tests whether the student's social environment and academic background influences his motivation and self-efficacy.

6th grade schoolchildren's motivation doesn't appear to be related to socio-cultural environment but the most socially advantaged students show a lower decline in motivation during lower secondary school. On the other hand, academic self-efficacy is linked to socio-cultural environment from 6th grade. From the beginning of lower secondary education, disadvantaged students experienced less academic self-efficacy.

16.30 – 17.00 Using Comparative Judgement to assess writing: too complex?
T. van Daal[1], M. Lesterhuis[1], V. Donche[1], S. De Maeyer[1]
[1]University of Antwerp, Belgium

Nowadays, comparative judgement (CJ) is used to assess writing. Judges compare two essays and decide which is of better writing quality. Evidence, however, indicates that judges experience some comparisons as too difficult. Consequently, they are more likely to take inaccurate decisions. As this provides a potential threat to the validity of CJ, research into CJ's complexity is highly needed.
This study conceptualizes CJ's complexity from an objective and experienced perspective. Objective complexity refers to comparison characteristics that enhance the amount of information to be processed. The quality difference between two essays is used as an indicator for objective complexity. Experienced complexity results from the interaction between the CJ task and a judge' ability and consequently varies between judges. Little is, however, known about antecedents of these differences in experienced complexity.
Therefor, this study focuses on the relation of training with judges' experienced complexity while taking into account quality difference and decision accuracy. Based on the theoretical framework, four hypotheses are formulated and their plausibility is examined using the CJ data of 14 judges that assessed 183 primary students' short essays. Judges were randomly assigned to either the training or control condition. Analyses are ongoing.

17.00 – 17.30 Crossing assessment cultures and overcoming the language barrier: The case of Syrian refugees and vulnerable youths in Lebanon
Y. El Masri[1]
[1]*University of Oxford, United Kingdom*

Young people moving to new countries face significant challenges in being successfully integrated into the educational system of the host country partly because of the distinct assessment cultures and practices of the host and home country. In Lebanon, Syrian refugees have been experiencing immense challenges to access quality education for various reasons, namely the language of instruction and assessment of mathematics and science being English or French and not Arabic (i.e. the native language of both Syrian and Lebanese people). A local non-governmental organisation developed an open access online facility that provides educational material in English, French and Arabic to minimise the impact of the language barrier.
This paper attempts to better understand what makes science tasks more difficult for deprived youth in Lebanon and whether the interactive nature of the tasks reduces the language barrier. Results of a study analysing a subset of science tasks available on the platform in terms of level of difficulty and demands using data from tests and cognitive interview will be discussed and recommendations for improving these tasks will be proposed.

16.00 – 17.30 Session DD: Issues in Comparative Judgement
Vltava and Vistula, Jon Šotola

16.00 – 16.30 The effect of adaptivity on the reliability coefficient in Comparative Judgement
S. Vitello[1], *T. Bramley*[1]
[1]*Cambridge Assessment, United Kingdom*

Comparative Judgement (CJ) is an increasingly widely used method for creating a scale, for example of the quality of essays. One popular approach for optimising the selection of pairs of objects for judgement is known as Adaptive Comparative Judgement (ACJ). It has been repeatedly claimed in the literature that ACJ produces very high reliability, often higher than can be obtained by conventional marking. It has been shown by simulation that adaptivity can substantially inflate the apparent reliability in ACJ. The aim of this study was to see if the same inflation would happen in real data by comparing an adaptive with a non-adaptive CJ study using English essays. An all-play-all set of comparisons of a subset of the essays allowed the extent of scale inflation to be quantified: the reported ACJ reliability was 0.97 whereas the all-play-all value was 0.82. The value from the non-adaptive study was 0.70. However, the scale from the non-adaptive study correlated slightly higher with external variables (the Principal Examiner's mark for the essay that was judged, and for a different essay), suggesting the non-adaptive study was no less valid than the adaptive one. The advantages and disadvantages of using adaptivity in CJ will be discussed.

16.30 – 17.00 Interpreting the validity of misfit statistics in Comparative Judgement
R. Bouwer[1], *S. Verhavert*[1], *M. Lesterhuis*[1], *R. Van Gasse*[1], *V. Donche*[1], *S. De Maeyer*[1]
[1]*University of Antwerp, Belgium*

Comparative Judgement (CJ) represents the shared consensus of multiple assessors, increasing both the reliability (Pollitt, 2012) and the validity of the judgements (Van Daal et al., 2016). However, as assessors are allowed to use their own conceptualization of quality, individual judgements can deviate from the group consensus. Misfit statistics can provide valuable insight into the extent individual assessors deviate from the group consensus (Whitehouse & Pollitt, 2012). Therefore, the aim of this study is to interpret the validity of

misfit statistics. Using CJ, 37 texts were judged reliably by 9 experienced assessors (SSR = 0.86). The same texts were judged by 16 less-experienced assessors. Misfit analyses indicated that of these assessors, four to seven deviated from the group, depending on the statistic being used. Additional analyses showed that misfits valued different aspects of text quality than the assessors who judged in accordance with the group consensus. From the assessors who were initially indicated as a misfit, only one persisted to deviate from the group after professional development. Hence, misfit statistics seem to validly indicate assessors who rate different aspects of text quality and they can be used to monitor and improve the quality of individual assessors.

17.00 – 17.30 From interesting theory to practical implementation: what we learned from piloting adaptive comparative judgement with a UK awarding organisation
A. Boyle[1]
[1]*AlphaPlus Consultancy Ltd., United Kingdom*

We report an evaluation of a pilot of adaptive comparative judgment (ACJ), and reflect on various claims that ACJ improves existing standards setting and maintaining approaches:

- Its supposed conceptual simplicity.
- The fact that ACJ does not require judges to envisage a minimally competent candidate.
- The argument that it is easier for human begins to make relative rather than absolute judgements.
- The alleged likelihood that ACJ will provide substantial gains in internal consistency reliability.
- We also posit that – if ACJ is to be considered as an improvement on current methods – then ACJ should correlate more closely with empirical outcomes than the previous standards setting method does.

We evaluate these claims using quantitative and qualitative evidence, and find that some are justified and others not.
Based on our research, we reflect on the proper functions of standards setting and maintaining, and what it means to provide stable and consistent measurement. We also reflect that introducing an innovation into a mature system where current approaches are well understood is not like operating in a vacuum. We should understand that existing approaches have quality, and that we need to show how an innovation can add value.

16.00 – 17.30 Session EE: Validity Issues in Test Development
Suite 1, Louise Hayward

16.00 – 16.30 Perception-based Evidence of Validity
T. Karelitz[1], *C. Secolsky*[2]
[1]*National Institute for Testing and Evaluation, Israel*
[2]*Mississippi Department of Education, United States*

With the evolving conception of validity, face validity was dismissed as a viable psychometric term. We revisit these ideas and conclude that although face validity should not be used to describe qualities of tests, studying how different constituents perceive properties of tests can be informative. Perceptions influence how a test is conceived, developed, implemented and evaluated. However, validity literature focuses mainly on score-based evidence (SBE), stemming from analyzing test performance. We propose that collecting and analyzing perception-based evidence (PBE) is useful for both test development and validation, specifically under the argument-based approach. PBE is derived from perceptions about various qualities of tests and scores, given by stakeholders such as test developers, content experts, test users, examinees, policy makers, and even lay public (whose opinion has recently become more pronounced by social media). At different stages of a test's life cycle, researchers can analyze PBE for different purposes: (a) to gain insights about items during test development and improvement, (b) to identify validity threats, (c) to enhance score-based validity evidence, (d) to generate alternative

arguments, and (e) to evaluate the clarity and plausibility of an interpretive argument. In conclusion, ignoring PBE hinders the ability to make compelling validity arguments.

16.30 – 17.00 Is knowledge familiarity a good predictor of item difficulty? Rethinking Webb's (2007) Depth of Knowledge scale

E. Sweiry[1], Y. El Masri[2]
[1]AQA, United Kingdom
[2]University of Oxford, United Kingdom

Predicting item difficulty is a challenge that is relevant to all assessment practices and remains considerable, despite the large number of variables identified in the literature as influencing item difficulty. Of particular interest is a variable we refer to as 'cognitive level', which typically ranges from knowledge or recall at the lowest level to skills such as synthesis at the highest. The variable is often measured using Webb's Depth of Knowledge (DOK) scale. Previous studies have, in general, failed to show any relationship between cognitive level and difficulty. This may be because items classified at the lowest level in DOK (i.e. recall) actually show considerable variation in difficulty.
This study, which uses science tests aimed at 11 year olds in England, is intended to investigate whether the use of separate rating scales for knowledge familiarity and higher-level skills would elicit a stronger relationship with difficulty than a single scale such as DOK. The paper addresses key findings and implications of the study, including the success of the scales in terms of the variance in difficulty explained, the extent of inter-rater consistency achieved, and the challenges in constructing a scale designed specifically to address the familiarity of knowledge.

17.00 – 17.30 Valid discrimination in the assessment of practical skills

S. Cadwallader[1], B. Cuff[1]
[1]Ofqual, United Kingdom

Practical skills in science are highly valued internationally, they transcend many cultural and linguistic differences between countries. Proficiency in a particular technique (e.g. titration) is sometimes conceptualised in terms of binary competence; the candidate is either competent in the technique or not. However, it may be that such an approach is too reductive, ignoring discriminatory information that could be used to differentiate candidates across a spectrum of proficiency.

To explore this issue further, fourteen examiners were recruited to assess video footage of 'mock' candidates undertaking practical activities in chemistry. A repeated measures design was used to compare the consistency of examiners' proficiency judgements across four different rating scales: 'fail/pass', 'fail/pass/merit', '1 to 5' and '1 to 10'. The dependent variable was the Gwet inter-rater reliability coefficient, which attempts to control for differences in chance agreement caused by differing scale lengths.

Findings suggest that examiners are just as reliable in discriminating between three grade levels (fail/pass/merit) as between two (fail/pass). However, examiners are less able to reliably apply five- and ten-point scales, suggesting that, even when effects of chance agreement are accounted for, these scales may be too granular.

19.00 – 22.30 Conference dinner *(Art Nouveau restaurant, Obecní dům nám. Republiky 5, Prague 1.)*

22.30 – 23.30 Underground dance *(Art Nouveau restaurant, Obecní dům nám. Republiky 5, Prague 1.)*

Saturday, 11th November

Symposia

9.30 – 10.30 **Assessment of Instructional Quality Across Cultures, Quantitative and Qualitative Studies. New Approaches and Findings**
Danube, Trude Nilsen

9.30 – 9.50 Bringing the content back into instructional quality research: Generalizability of generic and domain-specific observations
S. Blömeke[1]
[1]*UiO/ LEA/ CEMO, Norway*

A comprehensive observation protocol was developed that assessed the three established generic (classroom management, supportive climate, cognitive activation) and in addition two new domain-specific dimensions (mathematics and mathematics educational quality) of instructional quality (InQ) via direct in-situ classroom observations. Generalizability and decision studies were applied to 592 ratings to examine the stability or variability of InQ. 37 secondary mathematics teachers were observed 16 times by two randomly selected raters during two block periods à two lessons in 20 minutes intervals. Rater bias was low, items distinguished well between teachers. Generalizability and dependability coefficients were moderate to good. Most variance was attributable to teachers in case of classroom management, supportive climate, mathematics and mathematics educational quality. Cognitive activation differed because teacher performance varied more within block periods. Variability across block periods was particularly high in the case of mathematics quality. These differences indicate the need to re-conceptualize the concept of basic generic InQ dimensions. Sufficient reliability was typically achieved with one to three observers, with two to three lessons, and with two to five observations per lesson.

9.50 – 10.10 Using classroom videos and student surveys to measuring teaching quality
K. Klette[1]*, A. Roe*[1]*, M. Blikstad-Balas*[1]
[1]*University of Oslo, Norway*

Various measures are required when analyzing teaching quality and for the present study we combine classroom videos and student feedback measures. The study draws on video observations and student questionnaires from respectively 48 Norwegian language arts 8th grade classrooms (n=1125) and 49 mathematics 8th grade classrooms (n=1100), four lessons in each class, amounting to a total of 396 lessons. The video recordings were coded using the Protocol for Language Arts Teaching Observation (PLATO) that provided a systematic and validated protocol that resonates well with the conceptual frameworks for instructional quality outlined in this symposium. The Ferguson Tripod Survey, which also resonates well with key elements of instructional quality, was used to measure student perceptions and consisted of 38 items with five response alternatives.

Tentative findings suggest that in combination with systematic video analyses the student survey provides reliable and qualified information about the merits of different teaching practices. Both instruments may serve as diagnostic tools in the development of teaching at different levels.
Implementing specific procedures when collecting classroom data can increase trust in data and their results. This includes rigorous training and certification of observers and for the student survey – the assurance of student confidentiality.

10.10 – 10.30 Measurement issues and findings on instructional quality from PISA 2015 and TIMSS 2015

T. Nilsen[1]
[1]*University of Oslo, Norway*

The present study seeks to explore the reliability and validity of Instructional Quality (InQ) in the last cycles of Programme for International Student Assessment (PISA) and Trends In Mathematics and Science Study (TIMSS). More specifically the psychometric properties of InQ as well as its relation to student outcome are investigated.
Founded both on existing research on InQ and on PISA, a new scale for InQ was implemented as a national option in Norway, Belgium and Germany in TIMSS-2015.
Multi-level, multi-group Structural Equation Models were fitted and measurement invariance analyses were performed on data from PISA-2015 and TIMSS-2015 for Norway, Belgium and Germany.
Preliminary findings from TIMSS show overall good model fit for InQ, and the construct is metric invariant across the three countries, across grades, and across mathematics and science. For all three countries (TIMSS 2015, grade 8) and with the exception of cognitive activation, the relation between the factors of InQ and student mathematics achievement were positive and significant. The analysis of PISA data is in progress, however, in both PISA and TIMSS the findings point to the need for curve-linear approaches and for a need to establish a more subject specific scale for cognitive activation.

9.30 – 10.30 Assessment culture in Flanders – a story of added values
Amstel and Volga, Evelyn Goffin and Rianne Janssen

9.30 – 9.50 The added value of multiple methodologies in Flemish national assessments

J. Denis[1]*, E. Ameel*[1]
[1]*KU Leuven, Belgium*

In the process of data-analysis, we mainly use two methodologies: Item Response Theory (IRT) and multilevel models. This multiple methodology approach has major advantages, which will be illustrated with examples.
IRT is a theory of testing that puts test takers and items on one scale. This allows for setting a standard at the item side and infer the implications of that standard with respect to students passing the standard. IRT also allows for screening the measurement quality of the items. Another advantage is scale equivalence across measurement occasions, whereby Differential Item Functioning can be informative. Finally, with mixture IRT models we can make the distinction between quantitative and qualitative differences in mastery.
Using multilevel models, we investigate whether there are systematic differences among schools, classes and pupils in test performance. Subsequently, we examine whether student or school characteristics co-vary with these differences before and after taking other characteristics into account. The added value of this method is twofold. For the participating schools, the average performance of the school can be compared to that of similarly composed schools. For the research community, the multilevel models yield valuable information about which background characteristics are suitable to include in large-scale background questionnaires.

9.50 – 10.10 The added value of performance assessments in Flemish national assessments: the case of natural sciences

M. De Meyst[1]*, J. Denis*[1]*, S. Beringhs*[1]
[1]*KU Leuven, Belgium*

Performance assessments provide a valuable addition to traditional paper and pencil tests, especially to ensure construct validity for assessing certain skills and competences such as

communicating in a foreign language, conducting scientific experiments or solving technical problems. The combination of both allows researchers to draw up a more complete picture of student achievement.
However, the inclusion of performance tasks in a large-scale national assessments imposes severe challenges. Apart from the obvious organizational challenge to monitor individual student performance by test proctors in every school, there are a number of conceptual issues, ranging from translating curriculum standards into feasible tasks, rating test performance during or after the test administration, and setting a performance standard to ensure the generalizability of the results, to drawing policy conclusions. The presentation will focus on the rationale developed over the years to tackle these issues.
As an example, the different phases of a natural sciences performance assessment for students in the second grade of secondary education will be discussed in detail. Also, the costs and benefits of including performance assessment in large-scale national assessments will be reflected upon.

10.10 – 10.30 School feedback reports and parallel tests: adding (tangible) value to Flemish national assessments
E. Goffin[1]
[1]*KU Leuven, Belgium*

After the results of an assessment have been announced, each school from our sample receives a confidential feedback report about their own relative performance in the assessment.
We also release parallel versions of the assessment tests that remain available for several years. All schools can administer these 'parallel tests' themselves, following a standardization script, and send us the data. We provide them with a report similar to the one the assessment schools received.
Producing school feedback in general, and parallel tests in particular, constitutes a valuable expansion of our core task of system-level quality monitoring. It is factored in from the very onset of each assessment project. In this presentation we will touch upon some costs and benefits involved.
For schools, our feedback reports have become a respected tool for quality monitoring in their own right. Together with the commissioning educational authorities, who formally included the parallel tests in a recently launched tool kit for primary schools, we also put continuous effort into the promotion of their popularity. Feedback reports contribute to the visibility of our research centre in the field of Flemish education.

9.30 – 10.30 CAMAU Project: Progression Frameworks and Progression Steps
Vltava and Vistula, George MacBride

9.30 – 9.50 Assess what matters – Asesu beth sy'n bwysig
L. Hayward[1], J. Waters[2]
[1]*University of Glasgow, United Kingdom*
[2]*UWTSD, United Kingdom*

This paper outlines the theoretical underpinning to the CAMAU research. Since practitioners often currently lack a coherent framework of progression in learning in a curricular area to support their assessment practice, we argue that teachers require the support of explicit progression frameworks. These frameworks, rather than looking back and primarily recording what has been learned, should reconceptualise assessment as forward facing to identify the prerequisites for successful future learning. These frameworks therefore must identify 'what matters' in learning. 'What matters' has often concentrated on 'big ideas' or 'threshold concepts' which focus on knowledge; this must now extend to include skills, attributes and broad capabilities. Such frameworks as do exist are usually expert creations. Teachers commonly note that there is a mismatch between statements of standards or other curricular descriptions of

progression and the ways in which children and young people in fact learn. Progression frameworks which support valid assessment require therefore to be informed by empirical evidence as well as by research and policy developments in similar contexts. The paper concludes with consideration of how the assessment of 'what matters' can be used to support learners as they move forward to the next phase in their learning.

9.50 – 10.10 Subsidiarity and partnership – Sybsidiaredd a phartneriaeth

D. Morrison-Love[1]
[1]University of Glasgow, United Kingdom

The CAMAU project is committed in line with Welsh government policy to principles of subsidiarity and partnership. This paper provides a critical account of the development in this project of partnership working with practitioners to develop understanding of learning progression and support effective and sustainable policy and practice in progression and assessment. The paper identifies, outlines and analyses issues, theoretical and practical, related to extending the concept and practice of partnership in educational research. These include: the means of establishing initial relationships, developing collaborative practice and planning; making use of two languages of equal status. The paper provides a brief critical account of the Pioneer Schools networks established by the Welsh Government; of the collaboration of the CAMAU project with the networks; and of the relationship between CAMAU research related to assessment of learning and the networks' curriculum development work. At the heart of the collaboration between CAMAU and practitioners is the recognition of teachers as research collaborators. The paper affords a critical account of the development of research methodology and of differences between AoLEs, initial analyses of teacher produced data and consideration of means by which practitioners can be involved in critical consideration of research, policy and experience.

10.10 – 10.30 Learning about Progression – Dysgu am Ddilyniant

E. Spencer[1], N. Ryder[2] and S.V. Hughes[2]
[1]University of Glasgow, United Kingdom
[2]University of Wales Trinity Saint David, United Kingdom

This paper summarises the findings of two CAMAU research reviews: one of national policies related to progression; the other of research into models of progression in learning. After a brief description of the methodologies employed, the paper presents a summary of the key findings of the two research processes; we identify areas which require further consideration within the project. We proceed to identify and discuss critically implications of these findings for the development of progression frameworks focused learning, including both broad statements providing an overview of the journey from beginning learner to expert in a domain and detailed descriptions of progression in learning in topics within a given domain. Firstly we establish a set of principles which, we argue, should underpin any such development. Secondly we analyse a number of issues (conceptual and practical) identified as requiring to be addressed in any development of learning progression frameworks. Thirdly we introduce a decision tree structure which can be used at different levels to inform the development of policy which addresses the issues noted. Implications of our approach for the development of assessment policy and practice in other jurisdictions are noted.

9.30 – 10.30 Beyond classical statistics: different approaches to evaluating marking reliability
Suite 1, Ben Smith

9.30 – 9.50 Evaluating marking reliability using Generalisability theory
E. Harrison[1]
[1]*AQA, United Kingdom*

This presentation discusses how Generalisability theory (G-theory) (Brennan, 2001) can be used to obtain formulae for quantifying inter-marker reliability. The distinct feature of G-theory is its use of variance components: in essence, the error term in classical statistics is split into variance components that allow the user to determine which facet unreliability in marking can be attributed to. The method proposed for calculating inter-marker reliability is illustrated by applying it to data from the live monitoring of an AQA examination; it is also applicable to other UK awarding bodies' marking data, and to any marking monitoring data that captures multiple examiners' marks for a single response.
G-theory analysis provides item-level statistics on inter-marker reliability. The most valuable statistics are the reliability index (which takes values between 0 and 1, denoting how reliable the marking of an item is), and the standard error inherent in the marking. These statistics can be used to compare the reliability of disparate items, and to quantify how certain one can be that candidates have received their 'true score' for the item. The G-theory statistics are compared to those derived from a classical approach, demonstrating that they are comparable but more nuanced.

9.50 – 10.10 Evaluating marking reliability using the Many-Facet Rasch Model
W. Pointer[1]
[1]*AQA, United Kingdom*

The objective of the study presented is to analyse marker error using the Many-Facet Rasch Model (MFRM) developed by Linacre (1989). The MFRM is an extension of the rating scale model with an extra parameter that categorises rater (marker) effects. The presentation will illustrate the MFRM by applying it to data from the live monitoring of an AQA examination.
MFRM can be used to model a number of marker effects, including severity/leniency, inaccuracy, the halo effect and central tendency. As well as summarising at a group level it is also possible to look at individuals, which is useful in gaining a deeper understanding of any quality of marking issues. For example, the model can determine the spread of severity/leniency across examiners, but it will also allow for the identification of which examiners are the most severe/lenient. Another advantage of the MFRM is that it is able to handle missing data (i.e. not all candidates have to answer each question or be marked by every examiner) as long as there are no disconnected subsets. These occur when the marking design does not contain sufficient links between the levels of the facets involved (i.e. candidates, items and examiners).

10.10 – 10.30 Evaluating marking reliability using confirmatory factor analysis
Y. Bimpeh[1]
[1]*AQA, United Kingdom*

This presentation discusses the use of confirmatory factor analysis (CFA) (Bollen, 1989) to estimate marking reliability. G-theory is often touted as a conceptual centrepiece for evaluating marking reliability, along with Many-Facet Rasch models. However, neither the estimation method used in G-theory nor the classical methods of estimating reliability offer a clear way for testing violations of G-theory model assumptions or alternative factorial compositions of the true score; CFA can accomplish this.
In CFA, the marker score for each item can be expressed in terms of two components. The first component is the underlying 'true score', which is the common score shared between markers.

The second component estimates what is unique to a specific marker. The marker uniqueness component consists of two parts: namely, random marking error and systematic error. The marker systematic error contributes to unreliability regarding the true score. A CFA differentiates between the true score, the factor, and the attribute unique to each marker. A second order CFA can be used to resolve the confounding of random marking error and systematic error. The proposed method provides the model fit, which measures the extent to which the covariance predicted by the model corresponds to observed covariance in marking data.

10.30 – 11.30 Keynote Symposium
Validity Considerations for New Data in Performance Learning and Assessment
Convenors: Bryan Maddox, University of East Anglia and
Alina Von Davier, ACT Next, USA
Discussant: Paul Newton, Ofqual, United Kingdom
– Suite 1, Iasonas Lamprianou

10.30 – 10.50 Assessment and Validity In-Vivo
B. Maddox[1], B. Zumbo[2]
[1]UEA, United Kingdom
[2]University of British Columbia, Canada

In this paper we will consider the radical implications of 'in-vivo' ecological data from testing situations for validity conceptualisation, methods and design. We describe in-vivo perspectives as those that consider the dynamics, response processes and interaction of testing situations they occur in real-life settings – such as the household and the testing centre. Assessment in those 'noisy' ecological settings contrasts with the more sterile 'in-vitro' context of the laboratory. From the in-vivo perspective, validity design should be informed by the actual contexts, purposes and related consequences of test use. New assessment technologies are rapidly expanding the information base and richness of 'micro' and 'macro' process data to support improvised forms of analysis. As that data is integrated into 'dynamic' assessment design and the interpretation of test performance, this calls for new approaches and frameworks in validity practice. We will illustrate this argument with examples of such data, including 'micro-analytic' data on key strokes and mouse clicks, on interaction, talk, emotional affect and gesture, as well as larger-scale institutional, cultural and historical aspects of testing situations. We will discuss the implications for psychometric validity practice of this expanded information base, and suggest some radical conceptual implications for validity frameworks and rationales.

10.50 – 11.10 Computational Psychometrics and Validity Considerations for Multimodal Data
S. Khan[1], A. von Davier[2]
[1]ETS, United States
[2]ACT Next, United States

There is a growing need for assessment tools that capture a broad range of learner behavior necessary for the evaluation of skills such as problem solving, communication and collaboration. A key feature of such assessments is the use of interfaces that enable rich, immersive interactions and can capture multimodal process data i.e. a time series of multiple data mediums including audio, video and log files of student activity. However, the analysis of such data poses a significant challenge: how do we extract meaningful evidence of construct competency from complex performances as captured in varied and unstructured multimodal data? In addition, analyzing each of the multiple data modalities in isolation may result in incongruities and without appropriate use of context it may be difficult to interpret student activity as they show significant behavioral variations over time. To address these challenges,

we present a methodology that utilize advances in computational psychometrics and artificial intelligence. We propose to model the temporal dynamics and integration of multiple data modalities with a hierarchical analysis approach. This approach exploits concept hierarchies that reflect the nature of the data and goals of the assessment.

11.10 – 11.30 Performance learning and assessment within an argument-based approach
Saskia Wools[1]
[1]*Cito, Netherlands*

As the papers presented in this symposium highlight, innovations in learning science, technology and psychometrics provide us with opportunities to develop digital learning environments in which data on learning and assessment become entwined. Some of these environments even aim to support the learning of complex competencies. Data collected within these digital learning environments tend to be multi-modal in a sense that both process data as well as outcome data are used to make inferences about student learning. Although these inferences might differ from inferences in summative educational assessment contexts, it is still important to establish their validity for the intended purpose. This final paper explores the possibilities of using the argument-based approach to validation for digital learning environments with multi-modal data collection. It will discuss an interpretive argument that specifies inferences drawn within a learning context as opposed to inferences drawn in an assessment context. Furthermore, it will elaborate on combining data to establish a validity argument for these new inferences. The paper will conclude this symposium by reflecting on new threats to validity that arise in digital learning environments and when different data sources are combined.

11.30 – 11.45 Coffee

11.45 – 12.30 Keynote Presentation *(Suite 1)*
Chair: Thierry Rocher
Title: Assessing Assessment Cultures
Eckhard Klieme (German Institute for International Educational Research, DIPF, Germany)

12.30 – 13.00 Awards and Closing *(Suite 1)*
Thierry Rocher

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

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 2017 committee members are:

- Gill Stewart, SQA (United Kingdom) (Chair)
- Lesley Wiseman, Independent educational consultant (United Kingdom) (Special Interest Group)
- Amina Afif, Luxembourg Government (Luxembourg) (Newsletter Editor)
- Deborah Chetcuti, University of Malta (Malta) (Increasing Membership Approaches)
- Mary Richardson, UCL Institute of education (United Kingdom) (Social Media Manager)

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 2017 committee members are:

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

AEA-Europe | Prague Conference Organising Committee

- Andrej Novik (Scio, Czech Republic) (Co-Chair)
- Thierry Rocher (DEPP, France) (Co-Chair)
- George MacBride (University of Glasgow, United Kingdom)
- Jannette Elwood (Queens' University Belfast, United Kingdom)
- Guri Nortvedt (University of Oslo, Norway)
- Alex Scharaschkin (AQA, United Kingdom)

AEA Europe | Prague Conference Scientific Programme Committee

- Sarah Maughan (AlphaPlus, United Kingdom) (Co-Chair)
- Stuart Shaw (Cambridge Assessment, United Kingdom) (Co-Chair)
- Jana Straková (Institute for Development and Research in Education, Czech Republic)
- Radek Blažek (Czech School Inspectorate, Czech Republic)
- Cor Sluijter (Cito, Netherlands)

AEA Europe | Review Panel

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

- George MacBride, University of Glasgow, United Kingdom
- Angela Verschoor, Cito, Netherlands
- Rose Clesham, Pearson UK (corporate membership), United Kingdom
- Sandra Johnson, Assessment Europe, France
- Christina Wikstrom, Ůmea University, Sweden
- A Therese N Hopfenbeck, University of Oxford, United Kingdom
- Cor Sluijter, Cito, Netherlands
- Jana Strakova, Charles University, Czech Republic
- Andrej Novik, www.scio.cz, s.r.o, Czech Republic
- Dina Tsagari, University of Cyprus, Cyprus
- Alex Scharaschkin, AQA, United Kingdom
- Jannette Elwood, Queen's University Belfast, United Kingdom
- Guri A.Nortvedt, University of Oslo, Norway
- Maria Teresa Florez Petour, University of Chile, Chile
- Franciscus Kleintjes, Cito, Netherlands
- Ayesha Ahmed, University of Cambridge, United Kingdom
- Thierry Rocher, DEPP, France
- Gill Stewart, Scottish Qualifications Authority, United Kingdom
- Newman Burdett, n/a, United Kingdom
- Rolf Vegar Olsen, University of Oslo, Norway
- Iasonas Lamprianou, University of Cyprus, Cyprus
- Paul Newton, Ofqual, United Kingdom
- Sarah Maughan, AlphaPlus, United Kingdom
- Fabienne van der Kleij, Australian Catholic University, Australia
- Anton Beguin, Cito, Netherlands
- Stuart Shaw, Cambridge Assessment, United Kingdom
- Radek Blažek, Czech School Inspectorate, Czech Republic

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 2017 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 2017, the review panel were Elena Papanastasiou (Cyprus), Rose Clesham (United Kingdom) and Anton Beguin (Netherlands).

The 2017 Kathleen Tattersall New Researcher Award Winner is Fazilat Siddiq (Norway).

Notes

Notes

Sponsors

AQA
Realising potential

Cambridge Assessment

GradeMaker

Springer

Pearson

CITO

Partners

Vretta

AEA-Europe | Association for Educational Assessment - Europe

Assessment cultures in a globalised world

The 18th Annual AEA-Europe Conference