aea '21 EUROPE

Assessment for Changing Times:
Opportunities and Challenges

Conference Abstracts

“How to devise Visible Learning Assessment Capable Teachers and Students: Moving from ‘merely’ developing tests, to understanding interpretations from testing”

John Hatttie[1]

[1]University of Melbourne, Australia

Abstract:
So much of the current debates, training, and methods for teachers focus on psychometric properties such as reliability and validity. If, instead, the focus is on the adequacy and consequences of interpreting the reports from testing we would be in a much more successful state. This session explores score reporting, teaching students to interpret their test results, and argues that (for students and for teachers) unless there are consequential decisions, improvements, or actions from these reports then maybe we should seriously question the (over-)use of assessments in classrooms.

Bio:
Prof. Hattie is an internationally acclaimed researcher. He is Emeritus Laureate Professor at the Melbourne Graduate School of Education at the University of Melbourne, Chair of the Australian Institute of Teaching and School Leaders, and director of the Hattie Family Foundation. He has published and presented over 1600 papers, and supervised 200 theses students, and 60 books – including 28 on Visible Learning.
More information about his work can be found at:

https://findanexpert.unimelb.edu.au/profile/428067-john-hattie
https://www.visiblelearningplus.com/content/gold-papers
https://www.visiblelearningmetax.com

“Constructive grading to help teachers in their decision-making and to improve students’ learning: what else?“

Raphael Pasquini[1]

University of Teacher Education in the state of Vaud, Lausanne, Switzerland

Abstract:
For more than a century, grading has been one of the most hotly debated topics in education in almost all OECD education systems. Numerous studies have been conducted on that field. A worrying finding emerged: regardless of the level of education or the subjects involved, grading practices look like a hodge-podge and teachers lack the skills and knowledge to assess and grade consistently. Hence, it becomes urgent to think about education programs that aim to develop assessment and grading skills among teachers related to learning.
The conference will address this complex question. It will start by outlining some issues about grading. Then, a core definition of grading will be presented with several questions, in order to problematize the conditions under which grading practices can be constructive, i. e. related to learning and that can provide efficient feedbacks to students. We will next explain our theoretical framework and present a case study which will concretize the issues teachers’ have to manage when they shift their grading practices from an arithmetical to a constructive perspective. More global research findings and a definition of constructive grading will follow, before an open conclusion based on three challenges to be addressed.

Bio:
Dr. Pasquini, who is an Associate Professor in the Training and Research Unit in the Teaching, Learning and Assessment at the University of Teacher Education in the state of Vaud, Lausanne, Switzerland, will be presenting his keynote speech based on his dissertation in the online AEA-Europe conference 2021. The title of his keynote presentation is “Constructive grading to help teachers in their decision-making and to improve students’ learning: what else?”

“The True Impact Of Pisa On Education Reforms: Who Cares About The Evidence?”

Monserrat Gomendio[1]

Research Professor at the Spanish Research Council and co-founder of SkillsWEGO (consultancy)

Abstract:
International large-scale assessments (ILSAs) provide comparative evidence on how education systems perform and identify good practices which lead to better student outcomes. PISA was specifically designed to provide advice to policy makers and it is widely believed to have had a major impact on education reforms. However, PISA data show that after two decades student outcomes in OECD countries have not improved. The OECD acknowledges that its self-proclaimed mission has not been achieved but argues that it has successfully developed policy recommendations which have lowered the costs of education reforms; thus, it blames governments for failing to follow such good practices. I argue that the reasons why the evidence has not had any major impact are more complex. The evidence concerning the lack of impact of investment is strong, but the political costs of increasing class size and/or decreasing teacher salaries are huge due to vested interests. A second group of factors is strongly context-dependent, but the policy recommendations tend to be universal, leading to dire consequences. Finally, the evidence concerning variables that attempt to measure equity (a multi-dimensional concept) is partial and non-conclusive, so the policy recommendations have been heavily influenced by ideology. I conclude that the evidence provided by ILSAs interacts in such a way with vested interests and ideology that it does not make education reforms less difficult. I also examine for the first time the real impact of the available evidence on education reforms and show that it has been very small. This is partly to do with the quality of the evidence, partly with underlying conflicts of interest which play a much greater role than any objective data.

Bio:
Prof. Gomendio is a research Professor at the Spanish Research Council and co-founder of SkillsWEGO (consultancy). She started her career as a biologist and then moved into academia where she has held leadership positions as Director of the National Science Museum and Vice-President of the Spanish Research Council. She has also contributed to education reforms in Spain, in her role as Secretary of State for Education, Vocational Education and Training and Universities at the Spanish Ministry of Education, Culture and Sports (2012-2015).

She eventually joined the OECD (2015-2019) where she worked first as Deputy Director for Education and then as Head of the OECD Centre for Skills since it was created in June 2017. Her main role was to give advice to national governments on the policies that could be implemented to improve the level of knowledge and skills of the population, and make education and training systems more responsive to the rapidly changing demands from the labour market due to the impact of megatrends (digitalization, globalization and demographic trends).
More information about her work can be found at:
https://www.linkedin.com/in/montserrat-gomendio-0a6aba1a6/?originalSubdomain=es

Pre-conference workshops

9:00 - 16:30

How hard can it be? Issues around how best to provide evidence for assessment validity, reliability and fairness: the practice and challenge of validation

S. Shaw[1]

[1]University of Cambridge, United Kingdom

The responsibility for assessment providers to demonstrate robust and thorough validity evidence is a long-established expectation as are warnings about the “potentially serious consequences” (Kane, 2009, p.61) of shirking such responsibilities. Even assessment providers that have limited resources will still have a responsibility to demonstrate the quality and validity of their assessments. This workshop - intended to make the complexities of validation theory and practice less challenging and more readily operational, will consist of an introductory overview followed by five sessions (each punctuated by group discussions) taking as their focus: theoretical challenges; practical challenges; contemporary views of validation practice; construction of validation arguments; and the sufficiency and relevance of validity evidence. The final session will provide practical guidance for the validation of educational assessments, describing in detail ‘how to begin’, ‘how to proceed’, and ‘how to evaluate validity evidence’. By sharing experiences through a collaborative workshop environment, greater insights will be drawn leading to an increased understanding of the validation process and how it might be routinely operationalised in differing contexts.

Kane, M.T. (2009) ‘Validating the interpretations and uses of test scores’. In R.W. Lissitz (ed.), The Concept of Validity. Charlotte, NC: Information Age, 39-64.

Put your test to the test: Assessing test quality

B. Hemker[1], C. Sluijter[1]

[1]Cito, Netherlands

Educational tests serve a specific goal, such as evaluation, monitoring, diagnostics, selection or guidance. Such a goal is only met, if the test is of sufficient quality. This workshop aims to provide participants with practical tools to evaluate the quality of a test.
Our target audience consists of people involved in test development. Participants should have experience with at least some of the elements of test production. They also should have an understanding of the basic psychometric principles of testing. In the theoretical part of the workshop we give an overview of evaluation systems and show their similarities and differences.
In the practical part of the workshop we put the theory to practice, by having participants actually evaluating the quality of a test of their own choice, based on relevant information pertaining to the test. This includes possible research reports on how the norms are determined, and the reliability and validity of the test. The workshop leaders assist participants in applying the evaluation criteria to their own test.
In the final discussion, the findings of each participants are discussed and we round off with a list of practical lessons learned.

Item Banking and the Assembly of Test Forms

A. Verschoor[1], R. Visseren[1]

[1]Cito, Netherlands

The workshop starts with an introduction to item banking as part of the test development cycle, from the perspective of the test developer. The theory and some best-practices are presented: why is item banking an important issue; how can we make item banking be profitable? An overview on item banking systems will be given, and participants will be encouraged to share their views and experiences with item banks.
The participants will learn about the main features of the test construction process. We will practice specification of test requirements, when available from the participants themselves.
A short introduction in the use of Item Response Theory and Classical Test Theory in different types of test design will be given. Special attention will be paid to the development of multiple parallel test forms. The use of these multiple test forms will be discussed, as well as the requirements that must be fulfilled.
Also, various aspects of item bank maintenance and renewal will be discussed: how can we identify potential shortcomings in the available item pool, what role do security issues and item renewal schemes play in a project? Developing long-term views in item banking will be the main topic here.

Session A - Psychometrics I

13:45 - 14:15

Metaphors and the psychometric paradigm

T. Bramley[1]

[1]Cambridge Assessment, United Kingdom

A recent book (Baird et al., 2018) introduced the idea of different ‘paradigms’ for understanding the variety of standard-setting practices in educational assessment encountered around the world. This presentation explores the psychometric paradigm especially in terms of the underlying philosophical position(s) taken on the nature of psychological attributes, the definition of measurement and the conceptualisation of standards on tests of educational achievement. First I discuss the extent to which the concept of measurement is a metaphorical one when applied to educational and psychological attributes. I then take the position that the targets of educational and psychological measurement belong to the philosophical category of ‘powers’ and discuss what this means for our understanding of whether and how they can be measured. I then discuss the psychometric definition of a standards - that they are points on an abstract line that divide the line into segments to which we wish to apply verbal labels such as ‘distinction’, ‘merit’, ‘pass’ and ‘fail’. I conclude that the attraction of the psychometric paradigm is the conceptual clarity it brings through its use of the measurement metaphor, but also that when it comes to standard-setting it cannot solve all our problems.

Achieving parity of standards between paper-based and computer-based tests

L. Miller[1], S. Hughes[1]

[1]Pearson, United Kingdom

2020 highlighted the shortcomings of the existing system of traditional series-based, paper-based exams taking place in schools and centres and showed the need for more flexibility in assessment. Moving forward there is a need for true computer-based, on-demand tests that can be sat at home, not just paper behind glass. However, this brings the challenge of ensuring the standard set on the new computer-based test aligns to the existing standard on the traditional paper-based test.

In the qualification this research is based on, the construct of the two test modes is the same, but the test forms, some item types, raw marks, delivery and inherent standard setting methodologies are different.

The purpose of this will be to set the standards on the computer-based test that align to the existing standard set on the paper-based test.

This work would be of interest to assessment professionals working in test development, assessment design and standards and awarding functions who are looking to the near future and exploring the alignment of standards across a multi-modal qualification.

A Mixture Nominal Response Model with Response Certainty for Exploring Distractor-Related Misconceptions and Confidence in Answers

C. Chen[1], B. Andersson[1], J. Zhu[2]

[1]University of Oslo, Norway
[2]Hong Kong Baptist University, Hong Kong

Distractor-driven multiple-choice items were developed with misconception-related-distractors for diagnosing student's misconceptions. The certainty response indices (CRIs), which measure student confidence in answering items, were used together with distractor-driven items and provided evidence of student's confidence level when misconceptions exist. The presented study fulfilled a psychometrical research gap of developing an item response model together with CRIs to model the probability of selecting distractors and expected CRIs concurrently for exploring the relations between distractors, misconceptions, and CRIs. To demonstrate the model application in practice, a seven-items test measuring algebra was administrated by 773 grade-7 students. The seven items in the test contained three items not related to misconception (non-misconception items) and four items involved with the misconception of "collecting like terms with powers" (misconception-related items). The results showed that the proposed mixture nominal response model identified a group of students who had higher probability of endorsing distractors in answering items and lower expected confidence level in answering misconception-related items than non-misconception items. The implication of the finding was discussed.

Session B - Educational Policy

13:45 - 14:15

Reflections on the enactment of educational policies: The case of continuous assessment in Malta

M. Buhagiar[1], D. Chetcuti[2]

[1]University of Malta, Malta
[2]Faculty of Education, University of Malta, Malta

There are different ways of conceptualising and implementing continuous assessment within an educational system. In this presentation, continuous assessment refers to a formally planned programme, developed and implemented at national level, which is used for both formative and summative purposes. In this arguably formalised form of continuous assessment some of the assessment activities that take place within classrooms carry important consequences that serve managerial rather than professional purposes. Current efforts to introduce a national continuous assessment system in Malta provide an opportunity to reflect and draw insights on the often complex relationship between educational policies and practices. These insights would appeal in particular to individuals interested in the operations of educational systems that, like Malta, tend to struggle in policy enactment. The ongoing introduction of continuous assessment in Malta is treated in this presentation as a 'case' that is explored through comparative analysis between the documents produced by policy makers and the actions, not limited to assessment practices, that have ensued so far at school and classroom levels. This comparison, which is embedded within sociocultural understandings of assessment and education systems, has the potential to shed light on the intricate relationship between policy making and policy enactment.

How curriculum and assessment policies affect assessment culture in the English discipline: A New Zealand Case study

H. Fjørtoft[1], M.K. Lai[2], M. Li[2]

[1]NTNU Norwegian University of Science and Technology, Norway
[2]University of Auckland, New Zealand

This paper considers how New Zealand's assessment culture is shaped by curriculum and policy requirements and how this impacts English teachers' use of assessment results and other available data. New Zealand assessment policy strives for a balance between improvement, accountability, and sustainability purposes. We focus on the disciplinary specific needs of English teachers and their particular epistemology. We conducted a qualitative content analysis on 47 New Zealand policy documents, test-related materials, and peer reviewed research related to reading assessment in the New Zealand context. Data is broadly conceptualized, with both quantitative (e.g., test scores, attendance data, and student engagement data) and qualitative (e.g., observations, videoing, and students' and parents' perceptions) data available. National policies offer substantial support in reading, interpreting, and communicating reading literacy data. Furthermore, policy initiatives encouraging teachers to use results for improvement efforts are comprehensive, echoing efforts to balance accountability purposes with improvement and sustainability. However, both psychometric assessment tools and classroom assessment guidelines tend to emphasize generic over disciplinary specific reading literacy. This may negatively affect English teachers' use of assessment data. The lack of sensitivity towards English teachers' needs underlines the importance of taking disciplinary epistemologies into consideration when developing assessment policies.

The Future of qualifications and assessment in England: Stakeholder consultation outcomes for bottom-up reform

H. Dalton[1], L. Watts[1]

[1]Pearson Education, United Kingdom

As part of a ‘post-pandemic’ review of the qualifications and assessment system, a wide-reaching consultation took place involving over 6000 participants, probing views on the 14-19 phase education phase in England. These findings help to form a vision for what the next phase of educational reform in England should focus on.

The sample comprises over 5000 online surveys completed by: learners, former students, parents, teachers, and employers. In addition, an online public consultation attracted over 950 responses, 13 longer written or oral responses and 17 semi-structured interviews with individuals representing the spectrum of education expertise and policymakers. In addition, 150 Members of Parliament were polled for their views on the purposes of education and assessment.
This paper maps out how stakeholders see assessment and how this relates to the overarching purpose of education The data show the extent to which the various stakeholder groups value traditional forms of linear examination compared to other modes such as continuous, modular and teacher-marked.

The findings from this research offer perspectives on the extent to which bottom-up approaches to education reform, advocated by the OECD, and using diverse stakeholder views, can be implemented on a national level.

Session C - Test Development

13:45 - 14:15

How to Measure Efficiency of Hints in Interventionist Dynamic Assessment?

M. Skryabin[1]

[1]Independent Researcher, Russia

Dynamic assessment is an interactive approach to psychological or educational assessment when intervention embedded within the assessment procedure. It is based on Vygotsky's ideas about the zone of proximal development (ZPD) in support of the unity of learning and assessment.
There are two major approaches to dynamic assessment. The interactionist approach centers on an interactive and qualitative interpretation of ZPD. The interventionist approach focuses on the quantitative interpretation of ZPD and the following features of dynamic assessment can be identified: (1) a set of hints is determined in advance and offered to learners as they move through a test item by item; (2) the hints are arranged on a scale from implicit to explicit.
To measure the efficiency of hints, we used dynamic item response theory models where students' abilities change during assessment. In our model, the item parameters are not the only difficulty but the item learning effect that influences student ability if a student gave an incorrect answer and a hint was provided. We implemented Bayesian filtering and smoothing to dynamic generalized linear models to estimate item learning effects both for simulated and real-world data.

How do changes in the internal assessment system affect the results of external assessment of NIS students?

S. Yessenaliyeva[1], A. Shilibekova[1], Z. Jumabayeva[1], A. Zhapparova[1]

[1]Nazarbayev Intellectual schools, Kazakhstan

Nazarbayev Intellectual Schools (NIS) are a network of STEM schools that selects talented and motivated students.
This paper presents the results of a 5-year longitudinal research conducted in 20 NIS during the 2016–2020 period. An attempt has been made in this paper to assess the effect of changes made in the internal assessment of students' academic achievements in the secondary school level. The changes were based on introducing organisational solutions to ensure the integrity of the policy and practice of internal assessment, as well as its continuity with the external assessment as a significant validating indicator of the education quality. The paper describes, particularly, a new system of internal assessment implemented in 2016 and its impact on NIS' grade 10 students' academic achievement, particularly in mathematics, languages (Kazakh and Russian) and science.
The paper concludes that even though the assessment system has become more demanding (as evidenced by the participants' feedback), it was assessed positively by all. In turn, a positive relationship between the results of internal and external assessment allows formulating conclusions about its effectiveness for further use and development.

Reader’s cognitive actions as parameters of item difficulty: lessons for test developers

A. Ivanova[1], I. Antipkina[1]

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

Reading comprehension is an important skill which attracts attention of both education researchers and policy makers. Teaching to read is, to an extent, shaped by international comparative studies, including PIRLS. PIRLS reading framework has become popular in Russia. This study focuses on the analysis of an instrument developed specifically for the purpose of relating reader’s cognitive actions to the difficulty level of items. The instrument is based on the framework which combines PIRLS theoretical foundations and recommendations of International reading society. The assessment is aimed at 3d-graders and is administered online during one lesson. Two experts in teaching to read independently defined cognitive actions required by each item and then reconciliated their ideas with the third expert. The final list of cognitive actions was used to predict the difficulty of test items. Using a Linear Logistic Test Model approach, the effects of the actions were estimated. The application of the model in reading comprehension test validation is discussed.

Session E - Covid-19 Research I

13:45 - 14:15

Assessing Numerical Reasoning On-Screen Using Hint Items

D. Budzynski[1], M. Turner[1], B. Smith[1]

[1]AlphaPlus, United Kingdom

Numerical reasoning is typically assessed via posing scenario questions and using multi-mark item types, where partial credit can be given for correct working even if a computational error is made. This poses significant challenges for assessing numerical reasoning online.

AlphaPlus leads a consortium responsible for moving Welsh National Tests online; the final subject to switch format is numerical reasoning. This presentation discusses the process of developing and trialling “hint” items to assess numerical reasoning on-screen.

Hint items are a novel item type that uses a system of hints to scaffold learners towards the correct answer if they do not answer correctly on their first attempt. Despite the abrupt changes to teaching and learning because of the coronavirus, the data available broadly indicates the Numerical Reasoning Personalised Assessments are working as intended: supplying additional hints made the item easier in over 90% of cases. User views of this completely new item type were somewhat mixed; many teachers and learners valued the hint items, but there was some uncertainty about exactly how hint items worked.

The presentation will discuss the process of implementing this novel item type, and in particular the lessons learned about what works and what does not.

Language Assessment under COVID-19: What about teachers?

D. Tsagari[1], A. Maaoui[2], H. Dammak[2]

[1]Oslo Metropolitan University, Norway
[2]University of Tunis, Tunisia

With the global outbreak of COVID-19, educational institutions and teachers worldwide have encountered challenges in implementing assessment in different educational levels. The disruption of the pedagogical process has made it necessary for the majority of countries to take measures for pedagogical continuity online. Despite the sudden shift occurring in many settings to online teaching, little is known about teacher assessment practices and challenges during the crisis. In line with the newly emerging body of research on the move to remote teaching and assessment, this presentation explores the new assessment realities and requirements added during the COVID-19 period by examining the range of assessment measures and practices taken in different countries in primary, secondary and higher language education. An online survey was administered to teachers to scrutinize in addition to examining possible applications of online assessment. The presentation will discuss the findings as well as theoretical and practical implications for teachers' language assessment literacy in times of crises when the adoption of online assessment may become a necessity rather than an option.

Session F - E-Assessment I

13:45 - 14:15

E-portfolios in teaching, learning and assessment: Tensions in theory and praxis

E. Walland[1], S. Shaw[2]

[1]Cambridge Assessment, United Kingdom
[2]University of Cambridge, United Kingdom

E-portfolios are increasingly used in innovative ways, particularly in higher education, where they have the potential to transform teaching, learning and assessment. Given that students are learning in a hypertextual, digitalised and multimedia world, there is an ever-pressing need for assessment to be more authentic and engaging, and to further develop transversal skills. This research analysed the challenges and opportunities of e-portfolios in schools and universities. It aimed to investigate the extent to which e-portfolios can be used to assess skills such as collaboration and reflection in a fair, reliable and valid way, and to develop recommendations for implementation. A critical review of the literature on e-portfolio theory and praxis was conducted. 95 studies conducted across Europe (with a few notable studies from the US) over the last 20 years, were evaluated. Our analysis identified several tensions relating to e-portfolio theory and praxis, which arguably need to be resolved before e-portfolios can be successfully implemented. We found three main categories of tensions: 1) the underpinning theory and research; 2) the uses and purposes of e-portfolios; and, 3) the challenges and opportunities related to implementation. We propose internationally relevant recommendations on e-portfolio assessment and offer guidelines for implementation.

Towards a new framework of digital assessment items: linking item types to assessment objectives

C. Jongkamp[1], R. Hamer[2]

[1]Cito, Netherlands
[2]International Baccalaureate, Netherlands

Since the early days of digital assessment, test developers still mainly rely on variations of the constrained response item types. Whilst easy to auto mark and implement in testing, constrained item types are not necessarily the most appropriate instruments to assess complex thinking skills. To assess complex problem-solving strategies using authentic digital environments would require more advanced types, such as interactive responses, gaming, simulations and even augmented reality (e.g. Bressler & Bolton, 2013). However, in addition to reliable marking guidance and processes, empirical evidence linking item types to assessment objectives is often missing.
The International Baccalaureate (IB) and Cito are collaborating in developing a new and intuitively easy to use framework of digital assessment item types. This framework was used in a series of workshops to explore the possibility of creating the missing link between item type and assessment objective. In these workshops, assessment experts used a Revised Bloom's taxonomy (Heer, 2012) to link a set of digital exam items to assessment purposes. This paper presents results, showing that more constrained item types are mostly associated with factual and definitional knowledge and understanding, while assessment of complex thinking most often uses free response question types or uploaded digital content.

Researching accessibility in a high-stakes digital examination environment

I. Custodio[1], B. Maddox[2], D. McVeigh[3], N. Care[2]

[1]Pearson, United Kingdom
[2]UEA, United Kingdom
[3]Pearson Qualification Services, United Kingdom

This paper describes exploratory research that examines the user experiences and response processes of test takers with disabilities as they engage with digital GCSE exam questions. Assessment organisations in the UK have a legal and moral responsibility to deliver accessible examinations that are also valid, reliable, equitable and fair. Whilst digital assessments provide opportunities for technologically enabled adjustments to be built into item and platform design, it is not always apparent how that design can best meet the needs of test takers with disabilities that are diverse and scalar. Technical standards and accessibility guidelines such as WCAG 2.1 provide a useful framework, but it is vital that we also understand the user experience. This presentation therefore focuses on the experience of test takers with a range of disabilities and how design decisions and features impact on test and item accessibility. This is a complex area in which one size does not fit all. Goals of universal design have to be balanced with the specific needs and lived experience of individual learners with disabilities. We consider how these findings might inform a coordinated approach to inclusive assessment design.

Session G - Reliability

13:45 - 14:15

Evaluating the simplified pairs method of standard maintaining using comparative judgement

T. Benton[1], T. Gill[1]

[1]Cambridge Assessment, United Kingdom

Standard maintaining refers to any activity designed to ensure that it is no easier (or harder) to achieve a given grade or above in one year than in another. Various ways of using comparative judgement (CJ) to inform standard maintaining have been suggested in the past. This paper describes and evaluates a new method (simplified pairs) of using CJ in this context. The method is simple to implement as it does not require the estimation of script quality measures using a Bradley-Terry model.

We will explain how the new approach improves efficiency by allowing more exam scripts to be included in a study without increasing the amount of time needed from judges. We will then present the results from three experiments designed to test its accuracy in three examination subjects (English Literature, Mathematics and Science). In each experiment, the CJ approach to estimating the relative difficulty of two tests was compared to statistical equating based upon common students. In two of the experiments the results from the CJ method closely matched those from equating. However, in the final one (science) the results were less encouraging. We will discuss the possible reasons for the difference in results.

Comparing teacher judgement and exams: A case study of the factors influencing GCSE and A level grades in England in 2020

T. Stratton[1], N. Zanini[1]

[1]Ofqual, United Kingdom

In summer 2020, GCSE and A level exams were cancelled in England. Instead, schools allocated Centre Assessment Grades (CAGs) based on teacher judgement.
We utilised data on CAGs in 2020 and exam results from the previous two years, to identify if there were any changes in the patterns of relationships between grades awarded by teacher judgement in 2020 and other candidate, school and subject level features, when compared to those relationships in a 'normal year'.
Analysis showed that across both GCSE and A level there had been an overall increase in results by around half a grade in 2020 compared to previous years. There was also less variation in grades in 2020, overall and at the candidate and school level. However, the majority of relationships between grades and other features studied had not substantially changed.
We discuss additional key findings from the analysis on how the relationships between grades and candidate, school and subject level featured differed in 2020 compared to previous years, as well as the generalisability of these results to other instances of teacher judgement.

The impact of marking vs moderating IB coursework

B. Smith[1]

[1]AlphaPlus, United Kingdom

Like many other awarding organisations, the International Baccalaureate (IB) was unable to administer its Summer examination series in 2020 due to the Coronavirus pandemic. The decision was made to mark all coursework, so the IB would have had explicitly reviewed some work from all candidates (for subjects utilising coursework). This is in contrast to the moderation procedure for IB coursework conducted in a normal series.

AlphaPlus was commissioned by IB to examine the impact of marking rather than moderating coursework marks in 2020. Given that in 2020, examiner marks are available for every candidate's coursework, it was possible to model what 'would have happened' if moderation were used instead. This data thus provided a unique opportunity to compare the two validation methods, using a substantial dataset from over 8,000 centre/component combinations.

The presentation will discuss our results in full. Key findings include that, when centre marks are controlled for, examiner marking appeared to be slightly more lenient in 2020, suggesting that some element of the marking task examiners conducted in 2020 may have been subtly different to the typical moderation task. The analysis also implies that centres may 'over-compensate' for changes in their cohort's ability from year to year.

Session H - Formative Assessment I

15:45 - 16:15

Peer assessment of writing: a powerful formative tool for improved student learning

E. Meletiadou[1], E. Meletiadou[1]

[1]London South Bank University, United Kingdom

In the last few decades, researchers and educational authorities have expressed their concern for EFL students' poor writing performance and lack of motivation. Researchers indicate that peer assessment (PA) can support a better integration of teaching with assessment of progress in learning. Bearing this in mind, the current study employed a pre-test post-test quasi-experimental design and aimed to explore the effect of PA and TA on EFL students' writing performance; the impact of PA and TA on EFL students' writing quality as opposed to TA only; EFL students and teachers' attitudes towards PA. The study outcomes indicated that PA and TA can have a moderately positive impact on students' writing performance and a similarly significant impact on EFL students' writing quality. EFL teachers and students' attitudes towards PA of writing were positive and they both expressed their wish for multiple forms of assessment. In response to the need for more experimentation, the present study provides a PA implementation model for secondary school EFL writing classes enabling teachers to improve students' performance and motivation which in so far has been absent.

Assess@Learning - digital formative assessment in classrooms: stakeholders' views

K. Livingston[1], J. Elwood[2]

[1]University of Glasgow, United Kingdom
[2]Queen's University Belfast, United Kingdom

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

During 2021 Student Dialogue Labs and Country Dialogue Labs were convened online to provide opportunities for identified stakeholders to have in-depth dialogues about their views and opinions on digital formative assessment. Involving students in dialogue labs is an original feature of the study, and allows for students in each of the five European countries to discuss experiences of digital formative assessment which will ultimately inform European policy implementation in digital formative assessment.

This paper will present an initial analysis of the qualitative evidence from the Country and Student Dialogue Labs held in the early part of 2021. The paper will also share considerations of the use of qualitative research approaches to capture and identify unanticipated outcomes of implementing DFA as well as discuss methodological choices around dialogue labs for educational assessment policy research.

Session I - E-Assessment II

15:45 - 16:15

The use of remote invigilation – awarding organisation views on its introduction and impact

S. Cadwallader[1], D. Tonin[1]

[1]Ofqual, United Kingdom

Remote invigilation (RI), also known as remote proctoring, refers to the use of technology to supervise an assessment from a separate location to that where it is actually taking place. RI seeks to maintain the integrity of the assessment while allowing the learner the flexibility of being able to complete it from their home or place of work. RI's prominence had been increasing steadily in recent years and the Coronavirus (COVID-19) pandemic has significantly accelerated this growth. The adoption of RI presents both challenges and opportunities for assessment.

Ofqual, the regulator for qualifications, examinations and assessment in England, conducted seven case study focus group interviews with staff from a variety of awarding organisations, along with interviews with learners who had recent experience of being assessed under supervision from RI. The intention was to deepen our understanding of RI and how it is deployed and experienced, therefore informing our approach to regulation. Findings from the analysis of these interviews will be discussed under the following five themes: diversity of remote invigilation systems, infrastructural barriers, malpractice, learner experience, and developing capability.

Machine Scoring of Open-Ended Items

A. Verschoor[1], R. Visseren[1]

[1]Cito, Netherlands

Machine scoring is one of the key advantages of e-assessment. Yet, it is limited to closed-response items. The easiest-to-develop items, such as those for which the answer is a single sentence, are usually left to paper.
In this presentation we will take a closer look at the complexity of the problem, and we propose a method for machine scoring based on decision trees. Decision tree methods use a training set of responses that were scored manually.
We performed a study on the accuracy of the proposed method using responses on approximately 300 different items. In all cases, random samples from the responses were taken as training sets, while the remaining responses were used to evaluate the accuracy of the machine scores. Usually, training sets of 100 responses produced correlations between 0.65 and 0.95. At the same time, several response sets were rescored manually by different markers. Correlations between these two human markers were in the same order of magnitude for the vast majority of items.
The method we propose is certainly not the final solution to the problem of machine scoring. But it may be possible to use the machine scoring as a suggestion to aid the human marker.

Items in Technology-Based Assessments: Examining the use of multimedia stimuli with post-primary test-takers

P. Lehane[1]

[1]Centre for Assessment Research, Policy and Practice in Education (CARPE), Dublin City University, Ireland

Many countries are now deploying online testing solutions for their terminal post-primary exams. These Technology-Based Assessments (TBAs) use items that employ a broad array of dynamic or static stimuli (e.g. animations, text-image). Although it is assumed that these can make TBAs more authentic, their impact on test-taker performance and behaviour is still unknown. To determine the extent to which the use of different multimedia stimuli can affect test-taker performance, an experiment was conducted with 251 Irish post-primary students using an animated and text-image version of the same TBA of scientific literacy. Eye movement (n=32) and interview data (n=12) was also collected to determine how multimedia stimuli can affect test-taker attentional behaviour. The results indicated that, overall, there was no significant difference in test-taker performance when identical items used animated or text-image stimuli. However, items with dynamic stimuli often had higher discrimination indices indicating that these items were better at distinguishing between those with high and low levels of knowledge. Eye movement data also revealed that dynamic item stimuli encouraged longer average fixation durations on the response area of an item. The implications of these and other findings, as well as recommendations for policy, practice, and future research will be considered.

Session J - Fairness & Social Justice II

15:45 - 16:15

What is a fair vocational assessment?

S. Shaw[1], I. Nisbet[1]

[1]University of Cambridge, United Kingdom

Despite its late arrival on the assessment and measurement front, fairness has assumed a position of prominence to the extent that now many commentators acknowledge three ‘gods’ of assessment theory: validity, reliability, fairness. Considerable attention has focused on how the three concepts apply to the design, construction, administration and scoring of academic tests. By comparison, however, little consideration has been afforded to vocational qualifications and assessments especially from the standpoint of fairness. This presentation seeks to address two questions:

- Does fairness mean the same thing for academic and vocational qualifications?
- What are the current challenges to fairness of vocational qualifications/assessments?

The first part of the presentation distinguishes a number of senses of fairness that can be confused in discussions about assessment. The term ‘vocational’ is then unpacked in terms of how it applies to qualifications and assessments. In the second part, different approaches to vocational assessment and their implications for fairness are explored. Next, a small group of practitioner responses to open-ended questions about the (un)fairness of vocational qualifications are reported. Finally, senses of “fair” most relevant to vocational qualifications are discussed and recommendations for their design made.

Diversity and inclusion in GCSE and A Level History

J.M. Ryan[1], C. Balaban[1], V. Armstrong[1]

[1]AQA, United Kingdom

2020 was a year of change: change that the coronavirus pandemic created globally, and change in the narratives surrounding diversity, equality and inclusion that the Black Lives Matter and De-Colonising the Curriculum movements drove into the international spotlight. More questions are being raised concerning the inequalities present in educational assessment systems, and more inquiry made into the imbalances of perspectives and representations in national curricula.
This paper presents research conducted by researchers at AQA into the views held by teachers and students at five schools in England on the nature of diversity, inclusion and representativeness in the examination specification and curricular documents for GCSE and A Level History. Findings suggest that teachers and students are aware of a “dominant narrative” present in history curricula and assessment in the UK: that the voices of ruling powers and peoples are amplified while the voices of the oppressed and marginalised are muted. Students have challenged the predominance of these Anglo-centric and Western perspectives, calling for more representation of diverse peoples, cultures and traditions both in the United Kingdom and in other regions of the world. History students and teachers see the world has changed and want history curricula to change along with it.

Measuring socioeconomic status among middle school students: Does atypicality of responses matter for predicting academic achievement?

V. Morkevičius[1], R. Erentaitė[1], R. Vosylis[1,2], E. Melnikė[1], D. Sevalneva[1], S. Raižienė[1,3], B. Simonaiteinė[1]

[1]Kaunas University of Technology, Lithuania
[2]Mykolas Romeris University, Lithuania
[3]Vilnius University, Lithuania

Measurement of SES among school students often relies on the indicators of parental education, occupation, and income. However, this approach often results in large proportions of non-response and a lack of consistency in student answers. An alternative way is to use composite measures with multiple items indicating family wealth, cultural possessions, and home educational resources, which solves the problem of non-response. However, little is known about the quality of students' responses to such measures. The aim of our study was (1) to explore the dimensionality of middle school students' answers on a composite measure of SES, and (2) to assess the predictive utility of extracted dimensions. Using 2012-2016 data from the National Survey of Student Achievement (samples of 8th and 6th grade students, N1=2987; N2=1701; N3=2252) we conducted multiple correspondence analysis for exploring the dimensionality of responses to SES items (k=9). The results showed that at least two dimensions are required to acceptably explain the variance of SES items. Inspection of the dimensions revealed that the first reflected low vs. high SES, while the second reflected typical vs. atypical (incongruent) students' answers to SES items. The typicality dimension added substantial explanatory power to the prediction of reading and math scores.

Session K - Covid-19 Research II

15:45 - 16:15

Understanding learning loss during the pandemic – implications for assessment

G. Grima[1], J. Golding[2]

[1]Pearson UK, United Kingdom
[2]University College London Institute of Education, United Kingdom

The focus on easily measurable estimates of ‘learning loss’ via commercial assessments may be missing important aspects of learning loss that are harder to quantify and often less tractable to address.

Across our longitudinal (2016-21) primary and upper secondary mathematics studies in England, we found aspects of the curriculum such as routine facts and procedures are relatively easy to teach remotely (though still not easy to learn remotely): what have been marginalised in both remote and back-to-school-constrained learning are deep and connected conceptual learning and those mathematical processes and dispositions we value highly in recent curricular reforms: mathematical problem solving, reasoning and communication, and mathematical confidence, resilience, and growth mindset. These are harder to teach (and learn) and arguably, hardest to assess. Therefore, when reflecting on learning loss, these outcomes need to be considered, and evidenced, alongside those aspects of the intended curriculum that are easier to assess.

We present aspects of our findings. The reported losses might have a massive impact on future as well as current ‘recovery’ learning, but not be exposed by commercial assessments. Sharing work such as ours widens our discussion and adds value to our understanding of pandemic-related learning loss.

A first look at the impact of Covid in Wales' personalised assessments

M. Turner[1], B. Smith[1]

[1]AlphaPlus, United Kingdom

The Covid-19 pandemic has significantly disrupted schooling across the globe, including through total school closures and shifts to online learning. This has led to concerns about the impact of these disruptions on children's learning.
AlphaPlus leads a consortium responsible for moving the Welsh National Tests for learners aged 6-14 to an online adaptive model, underpinned by a common IRT scale. The Procedural Numeracy Personalised Assessment (PPA) went live in 2018. The availability of historic data from the PPA, coupled with the use of a common IRT scale throughout each school year assessed, means that Wales is uniquely placed to be able to quantify any possible impact of Covid at cohort level.
We matched learners that sat the PPA in Autumn 2020 to their historic data. We found that pre-Covid, progress was not 'gained' equally across each month. However, by using pre-Covid data from a similar time in the year we were able to control for this. We found that at the whole cohort level, whilst some learners did make 'better than average' progress, fewer than the half of learners we would expect to did. There were also indications that disadvantaged learners may have been more affected by the pandemic.

Non-examined high stakes assessments in vocational education: identifying and managing risk post Covid-19

R. Bagguley[1]

[1]Pearson, United Kingdom

There is a desire to strengthen the approach to quality assurance for non-examined assessments to ensure that the results issued are defensible and instil public confidence. The on-going disruption to teaching and learning and introduction of new regulatory conditions in the UK for vocational qualifications from September 2021 highlight the importance of exam boards reviewing their current approach to quality assurance. Risks that would potentially undermine the credibility of the qualification need to be identified prior to quality assurance activities being conducted by exam boards, to provide confidence in the standard of non-examined assessments. Four common approaches to moderation in non-examined assessments are used with vocational qualifications internationally. Recent global crises may provide an insight for the education sector into effective risk management strategies as the introduction of new regulatory conditions and making adaptations to comply with these regulations is similar in many respects to the circumstances faced by the finance sector following the 2008 financial crisis. This session will explore the attributes of organisations who have been successful when responding to recent crises and how exam boards can respond to the challenges faced in a post Covid-19 context by embedding these behaviours into their approaches to quality assurance.

Session L - Formative Assessment II

15:45 - 16:15

Assessing, mapping, measuring and facilitating the development of abstract skills: how do teachers facilitate creativity and curiosity in classrooms?

T.N. Hopfenbeck[1, 2], J. Scott-Barrett[1], S. Johnston[1], T. Calabrese[3], J. McGrane[1]

[1]OUCEA, Department of Education, University of Oxford, United Kingdom
[2]Department of Teacher Education, Norwegian University of Science and Technology, Norway
[3]OUCEA, Department of Education, Oxford, United Kingdom

Creativity and curiosity are essential skills for life-long learning and are the focus of many recent developments and debates in educational and assessment practices and policies. However, how do teachers facilitate and assess the development of these skills? And how do researchers explore and observe these in classroom contexts? This paper explores the methodological implications of exploring how teachers facilitate curiosity and creativity across international contexts: we explored teaching practices through video observation, teacher and student interviews, as well as integrating two short creativity and curiosity activities to develop methodological instruments and age-appropriate scales to better understand the development of these skills among primary school children. This research was conducted by a research team from the Oxford University Centre for Educational Assessment (OUCEA) across six schools in collaboration with the International Baccalaureate and the Australian Council for Educational Research and funded by a grant from the Jacobs Foundation.

Decomposition of the Composite: Relating Students' Abilities from Item Response and Cognitive Diagnostic Models

D. Federiakin[1]

[1]HSE University, Russia

Cognitive Diagnostic Models (CDMs) allow students' multidimensional classification in terms of mastery or non-mastery of every sub-content area. Item Response models report some unidimensional factor, which allows to rank students in terms of their overall ability to solve particular test items. The difference in the models' ideology and purpose leads to usual mutual exclusiveness in their application: if one is used, the other is not. However, some tests allow retrofitting one type of the models, even if the test was developed in another paradigm. One of such tests is the Test of Understanding College Economics (TUCE). This test is built using a taxonomy of items in cognitive levels and intensive classification in sub-content areas. As a result, the test possesses a detailed cognitive map of the test items, which allows the calibration of CDMs. The test fits the unidimensional Item Response model very well, allowing us to report only one test score. As a result, we have two types of ability estimates: unidimensional IRT score and students' multidimensional classification from CDM. In this presentation, we describe and compare how these scores complement and relate to each other.

Implications of education reform on assessment practice in schools: Teacher and student perspectives on the enabling and disenabling characteristics of formative assessment.

J. Kellough[1], D. Murchan[2]

[1]Diocese of Kalamazoo, United States
[2]Trinity College Dublin, Ireland

The literature on assessment reform highlights the centrality of formative assessment (FA) to student learning. Given that FA is emphasised strongly in Ireland's recent reform of Junior Cycle, an investigation into the value-added by the new assessment initiative, the efficacy of its implementation, and its impact on stakeholders is imperative. Questions remain around the policy's purpose, its alignment with best practice, teachers' engagement/understanding of it, and the impact on pedagogy and student learning. Answering these questions has significant relevance for students and teachers in lower secondary education and for the reform of Senior Cycle currently underway.

This evaluative case study engaged teachers in professional development aimed towards promoting teachers' assessment literacy. Teachers and students reflected on the enabling and disenabling characteristics of specific features of FA: Learning Intentions, Success Criteria, Formative Feedback, Effective Questioning, and Self-Assessment. Themes identified from the research included promoting student engagement and ensuring FA's interdependence. Findings have implications for education systems attempting to introduce a structured approach to FA, especially where a culture of high-stakes assessment operates at the same educational level to shape teachers' perspectives.

Poster presentations

11:30 - 12:45

Developing e-assessment instruments for assessing 5-year-old children’s skills

A. Meesak[1, 2]

[1]Education and Youth Board, Estonia
[2]Tallinn University, Estonia

It is increasingly recognised that early learning is critical for children's development and affects later life. In 2019 we began developing innovative e-assessment instruments for assessing 5-year-old children’s development. An assessment framework and a test prototype were created together with experts in psychology, special and early childhood education. Children’s development is directly assessed in five key areas: cognitive, learning and social skills, literacy and numeracy. The instrument includes three thematic tests with different domains integrated into a continuous story. All items include audio instructions and are computer-assessed.
The purpose of the assessment instrument is to provide teachers with formative feedback on every child’s development and to get data to improve the quality of early childhood education. The tests are accessible in Estonian national examination system (EIS) and are specially developed to be child-friendly, use the advantages of e-assessment and work on tablets. An item trial for the first version of the two tests was carried out in 11 kindergartens in autumn 2020, and a trial for the third test will be carried out in 10 kindergartens in spring 2021. The poster will present our experiences and best practices derived from the e-assessment instruments’ development process and the first item trials.

The blind side: Exploring item variance in PISA 2018 cognitive domains

K. Marcq[1], J. Braeken[1]

[1]University of Oslo / Centre for Educational Measurement, Norway

Research on the Programme for International Student Assessment (PISA) focuses predominantly on response differences between pupils in their broader school contexts, whereas the investigation of the response differences between items rarely exceeds the original piloting, IRT scaling, and computation of the plausible values. In assigning such superiority to the analysis of test takers' performance, a void is created as that of the items is overlooked.

Viewing the variance in a test response as a combination of pupil, school, and item variation and given the broadness of the PISA constructs of reading, mathematics, and science, we hypothesize that the items' contribution to the total response variance is greater than that of the well-defined population of pupils and schools. If true, further inquiry into the item variance and the factors affecting its magnitude is instrumental in promoting a more comprehensive item-level validation framework and providing better diagnostic tools for future test development.

We use a random person random item modelling approach through a cross-classified mixed effects model to decompose the variance in the PISA 2018 cognitive domains response data. Variance components for pupils, schools, and items are computed to compare their relative contributions to the total variance structure.

Well-being and Academic Achievement of NIS Students: Improving Assessment Procedures

G. Sultanova[1], A. Rakhimbekova[1]

[1]Nazarbayev Intellectual School, Kazakhstan

The aim of this study is to find out how assessment procedures can be improved so that they have a positive impact on well-being and academic achievement of Nazarbayev Intellectual School (NIS) students. NIS is an experimental platform for the development, testing, implementation, monitoring of modern models of educational programs for secondary education levels. We hypothesize that cognitive load of NIS students and their satisfaction with assessment predict their well-being; well-being of NIS students, in turn, predicts their academic achievement. We also suppose that cognitive load and satisfaction of NIS students with summative assessment directly predict their academic achievement. Academic achievement is a latent construct comprising marks of internal and external assessment for a term. Cognitive load, satisfaction with assessment, and well-being of NIS students are latent constructs; each construct comprises a set of observable variables that will be gained via questionnaires. The sample consists of 10th grade (N=1729) and 12th grade (N=2295) students; the sample size is 4024 (26.37% of the total number). The data analysis is conducted by structural equation modeling. Studying the links between academic achievement, well-being, cognitive load, and student satisfaction with assessment can give us a first impression on how to improve assessment procedures.

Remote marking of high-stakes examinations: leadership, challenges and strategies

E. Walland[1]

[1]Cambridge Assessment, United Kingdom

The marking of high-stakes examinations has shifted over time such that activities and meetings are increasingly carried out remotely. This remote context provides various practical benefits, but it remains fundamental to ensure that students' work is marked to a high standard often within fairly tight deadlines. This qualitative study focused upon the marking of high-stakes examinations taken by secondary school students in England. Marking is carried out by a group of markers, led by a team leader, and this research investigated the qualities and behaviours that are perceived to be important for leadership in this context. Any potential challenges that take place when marking within a remote context were explored, alongside the strategies used to overcome them.

17 semi-structured in-depth interviews were conducted with assessment specialists and markers, who represented a wide range of different subject areas. A key finding applicable to all subjects was that leadership was focused around the shared goal of achieving fairness to students. Successful leaders were perceived to create a positive team culture and facilitate high quality marking through various strategies despite the lack of face-to-face interaction. The findings facilitate a nuanced understanding of the leadership of marking processes in a remote marking context.

Let the boys speak: stories of 12-year-olds about assessment

D. Said Pace[1, 2]

[1]Ministry for Education, Malta
[2]Institute for Education, Malta

Effective formative assessment (FA) is about the meaningful interaction that goes in the learning environment between the main stakeholders - students and teachers. However, the quality of this dialogue depends very much on the users' understanding and awareness of their role within FA. Several studies have been consistent in their findings about the assessment literacy of the teachers as low levels in assessment literacy correlate, amongst other factors, to a teacher-centred approach. Opting for such an approach could be the lack of belief in the students' competences and capability of being good FA users, thus limiting their involvement whilst adding the burden on the teacher. Establishing a shareholding partnership with students entails that the latter are equipped with the skills for handling FA. Positive pressure on the teachers can be through a bottom-up approach by analysing where the students are in their knowledge and skills of FA thus highlighting where teachers need to invest to empower their students. This is precisely what this study will try to do - listen to eight boys' voices about their assessment experience to share their narratives. Focus on boys was due to convenience sampling in also being a mother of two boys.

Random responders in international large-scale assessment in education: a Threat to validity?

S. van Laar[1], J. Braeken[1]

[1]CEMO: Centre for Educational Measurement at the University of Oslo, Norway

Data collected by international large-scale assessments provide us with a wide variety of research opportunities. Yet, international large-scale assessments in education are also typically low-stakes for the students, which makes the assessments susceptible to invalid response behavior. Consequently, students' responses might no longer reflect true knowledge, abilities or opinions related to the assessment content. Depending on the severity of this invalid response behavior, this could lead to problems with the interpretation of assessment results.

The aim of our study is to investigate the prevalence of random responders and their impact on inferences related to the TIMSS 2015 student questionnaire. To this end, we used a mixture IRT approach to identify those students who would qualify as random responder (i.e., students providing uncorrelated, or random, responses to the questionnaire).

To investigate the impact of random responders on different scale-related inferences, we will be comparing analysis results with and without random responders. We will present results for a selected sample of countries with respect to the Confidence in Mathematics, Value of Mathematics, Confidence in Science, and Value of Science scales.

Adopting a multilingual approach towards comprehension in assessment in Higher Education Institutions in the UK amidst the Covid-19 pandemic

E. Meletiadou[1]

[1]London South Bank University, United Kingdom

The multilingual approach towards comprehension in assessment focuses on the presumption that multilingual learners may face incredible challenges when they are assessed through the English language which is their second or even third language. De Backer, Van Avermaet and Slembrouck (2016) and Menken and Shohamy (2015) also stress the challenges of assessing content using exams with instructions in the target language. In terms of the current study, 100 foundation year students from different ethnic and linguistic backgrounds at London South Bank University Business School were encouraged to translate the instructions for their assignment and use their mother tongue to discuss their assignments with fellow students coming from a similar background. This is a way of working across languages by using one language for input and another for output which was also confirmed by a number of studies which have used translanguaging (Lewis, Jones and Baker, 2012). The current study has indicated that the use of translanguaging can increase students' writing performance and attitudes towards learning as well as support their well-being in HEI amidst the Covid-19 pandemic.

A comparison of outcomes from tests proctored locally in testing centres and online using live remote proctoring (LRP).

G. Cherry[1], M. O'Leary[1], L. Kuan[2], O. Naumenko[2], L. Waters[2]

[1]Dublin City University, Ireland
[2]Prometric, United States

This research examined the psychometric equivalence of a set of professional licensure examinations using an onsite proctor or taken online using a remote proctor (who is proctoring in real time i.e. live remote proctoring (LRP)). The use of remote proctoring had been growing for over a decade, particularly in higher education contexts. However, demand across professional licensing and credentialling programmes increased significantly with the onset of Covid-19. In order to continue testing, many organisations turned to remote proctoring. Due to its relative novelty, remote proctoring is currently under-researched. The present study was undertaken as a response to the dearth of empirical evidence to guide decision making about the comparability of test centre and remote proctoring modes.

Analyses focused on four psychometric issues of key interest to accrediting agencies: reliability, decision consistency, average item difficulty and item discrimination. In terms of overall candidate performance (n = 15,000), passing rates and psychometrics, this paper provides evidence that outcomes from test centres and remotely proctored tests are equivalent across 16 insurance credentialing exams taken in 5 US states. The results obtained from this study provide no empirical evidence prohibiting the use of live remote proctoring in the context of high-stakes professional licensure exams.

Session M - International Surveys

9:00 - 9:30

Cross-country differences in rapid guessing behavior in PISA 2015

M. Michaelides[1], M. Ivanova[1]

[1]University of Cyprus, Cyprus

Examinees' test-taking effort in achievement tests has been found to have significant impact both on their performance and on the psychometric properties of the test. Low-stakes assessment programs, in particular, pose few or no consequences for the test-takers, but greater consequences for the participating institutions. When test-takers do not invest adequate effort, test scores underestimate the individual's true ability; ignoring the impact of test-taking effort may harm the validity of test outcomes. The study aims to examine the level of examinees' test-taking effort and its relationship with their test performance in the Program for International Student Assessment (PISA) across countries. The 2015 PISA assessment in science, mathematics, and reading in over 50 nations was computer-based and provides a response time measure at the item level, which was used as a behavioral measure of students' test-taking effort. Secondary analysis on PISA data describes test-taking effort across nations and other demographic variables and its relationship to achievement. Examining the relationship between effort and performance in different groups of examinees can provide a better understanding of the phenomenon and enhance group score comparisons. It will also allow for different ways of dealing with the low test-taking effort for the different groups of examinees.

Using PISA data to examine high-achieving students' characteristics

V. Pitsia[1]

[1]Dublin City University, Ireland

In spite of policymakers' heightened interest in STEM in Ireland and elsewhere, evidence suggests that high achievers' needs are not being met by education systems. As a consequence, students may not be reaching their potential in mathematics and science. To assist the Irish education system towards this direction, this study investigates high mathematics and science achievement using Ireland's data from the Programme for International Student Assessment (PISA). Ireland provides an interesting context because, despite students' good average mathematics and science performance on national and international assessments, there is a notable absence of high achievers. The results indicated that students' self-beliefs, learning dispositions, engagement, and socio-economic background significantly predicted high achievement in these subjects. The multilevel models explained a considerable proportion of the variance in mathematics and science high achievement. Discrepancies between the study findings and those on average achievement prompt investigation of student and school characteristics at different performance levels to afford students of varying abilities the opportunity to achieve their potential.

Who Benefits from Improved Outcomes in Reading Literacy in Ireland? An Investigation of Equality Using National and International Assessment Data

A. Karakolidis[1], A. Duggan[1], J. Kiniry[1], G. Shiel[2]

[1]Educational Research Centre, Ireland
[2]Educational Research Centre, Dublin, Ireland

Large-scale assessments can have an important role in evaluating educational reforms, indicating patterns in student performance over time. However, performance improvements are not always evenly distributed, and may exacerbate performance gaps between subgroups.
Using multiple cycles of national (NAMER) and international (PIRLS) assessment data, this study investigates whether improvements in Irish primary school students' reading performance are accompanied by greater equality and alleviation of subgroup differences, after the introduction of a National Literacy Strategy in 2011. The study examines reading performance gaps based on selected demographic and socioeconomic factors over time. Also, multilevel regression models were constructed to compare variance explained by school and pupil-level factors across cycles.
The statistical analysis results indicated that after the introduction and initial implementation of the Strategy, subgroup differences in reading performances shrank and overall variance attributed to background variables decreased by 14.0% and 18.2% in NAMER and PIRLS, respectively. Additionally, there was a considerable decrease in variance in reading performance attributed to between-school differences over time for NAMER; this was not the case for PIRLS. Overall, these findings suggest that Ireland has made reasonable progress in addressing inequalities in education.

Session N - Assessing Mathematics

9:00 - 9:30

I know what to do!: Advantages and issues related to developing and administering digital numeracy assessments for 6-year-old students

G.A. Nortvedt[1], O. Kovpanets[1], A. Pettersen[1], A. Rohatgi[1], L. Øygarden[1]

[1]University of Oslo, Norway

In Norway, the national government provides primary schools with assessment tools to detect students at risk of lagging behind in basic skills. Traditionally, these assessments have been paper based, and students have taken the assessments with teacher support. Starting in 2022, the assessments will be digital. The aim of this paper is to discuss the advantages and issues related to developing and administering digital numeracy assessments for 6-year-old (or grade 1) students. One advantage of digital assessments for first graders who do not read well enough yet to understand written task instructions: digital assessment tasks can include audio files or visual aids that provide students with necessary instructions. In addition, the digital format allows for interactive tasks, enabling students to work in much the same way they do in classroom activities during mathematics instruction. At the same time, digital tests developed to assess on students' numeracy competence, can also put high demand on their listening ability, concentration and motor skills. Data from cognitive labs and pilot studies with large sample sizes will be used to discuss how students are affected, and how valid and reliable digital numeracy assessments can be developed for young students.

A study of gender, self-perception, and mathematics: The 2020 England, Wales, and Northern Ireland PISA Field Trial

M. Hill[1], G. Grima[1], I. Custodio[1], J. Golding[2]

[1]Pearson UK, United Kingdom
[2]University College London Institute of Education, United Kingdom

The divide between female and male participation in mathematics is not a new concern for many countries – dozens of papers have been written on the topic in just the last decade. Nor is this concern surprising. Normatively it is a question of social justice. If females are being disproportionately excluded due to conscious or unconscious decisions, these need to be addressed. Pragmatically it is an economic issue, as the failure to engage females in STEM subjects may hurt national economic advantage. While there has been some good news in the UK in recent years – more females than males took STEM A-Levels in 2019 – issues remain specifically with mathematics, where females continue to enrol at substantially lower rates than males.

This paper offers an overview of the current state of this issue in England, Wales, and Northern Ireland, making use of the 2020 PISA Field Trial. It does this, first, by identifying key issues as recognised in previous literature (prior attainment; enjoyment; perceived competence; interest; and perceived utility); and second, locating these issues within the 2020 dataset. By identifying the continuation and extent of these challenges, the paper has potential also to identify opportunities to address them.

Teachers Beliefs of Growth Mindsets and Students' Mathematical Skills

E. Abdurakhmanova[1]

[1]Higher School of Economics, Russia

The current study examine whether teachers' beliefs about growth mindset and efficacy beliefs predict the growth of mathematical skills in elementary school students. Children were given a START test at the start and at the end of 1st grade and PROGRESS test at the start of 3rd grade. The teachers were given an online survey at the start of 3rd grade to determine their beliefs about the nature of intelligence (mindset) and teaching efficacy. We try to answer the following research questions:
(1) Do teachers' beliefs about growth mindset predict their students' 3rd-grade mathematical scores, controlling for 1st-grade scores?
(2) Is there a relationship between teachers' beliefs about growth mindset and children's mathematical scores?
(3) Do teachers' efficacy beliefs mediate the relationship between teachers' teachers' beliefs about growth mindset and children's mathematical scores?
The first results demonstrate there is a relationship between teachers' growth mindsets and children's mathematical scores. However, the research has not been finished yet.

Session O - Covid-19 and Summative Assessment

9:00 - 9:30

Qualification results in England in summer 2020: Teacher judgement in action

E. Howard[1], S. Holmes[1]

[1]Ofqual, United Kingdom

Final 2020 assessments in England were cancelled due to the Covid-19 pandemic and qualifications were awarded based upon teacher judgements of the grade students would have achieved had they sat the assessments. To understand the process used, and the views of those involved, we ran an online survey, for which we received over 1,200 responses, and conducted in-depth interviews with 54 of those survey respondents.

Many of the centres ensured consistency of judgements by centrally-designing the process to be used by individual departments. Usually the process was data-led, both to ensure objective judgements were made, and to ease evaluation of all of the different sources of evidence. Mocks were usually highly weighted, but other sources of evidence were used. Some judgment of student trajectory was often allowed, although respondents reported the difficulty in making evidence-based judgements for the likely performance in a final assessment for those students who might have relied on last-minute revision.

Confidence was high in the judgements overall, and there was a strong belief that bias had been minimised, but the entire process had been stressful for many. We will also reflect on some differences coming to light from a provisional analysis of the summer 2021 process.

Reshaping external high-stake assessment to mitigate students' inequality during the first stage of Covid-19 pandemic in Portugal

G. Cipriano[1], S. da Cruz Martins[1]

[1]CIES_ISCTE, Portugal

Worldwide, the covid-19 pandemic brought significant changes to teaching and learning processes, with impacts on students' learning and educational inequality. Seeking to mitigate the impact of external high-stake assessment on students' progression and grade transition in Portugal, towards the conclusion of the 2019/2020 school year, an exceptional consensus among educational communities has emerged, allowing a fast implementation of relevant educational assessment policies by the government, schools and teachers. This paper aims to explore these educational assessment policies implemented in Portugal during the first stage of the covid-19 pandemic, from March until July 2020. The main changes of assessment processes are discussed, highlighting their collaborative and multi-level decision-making characteristics, and the impacts that such changes had on students' outcomes. Preliminary institutional data suggests that, though students' learning has worsened, those changes were able to fulfil their goals, creating a positive impact on grade transition and students' progression, and thus reducing educational inequality. Despite the existence of a positive impact on students' outputs, some of those changes are unlikely to remain in the future due to validity and reliability issues. Nevertheless, some might remain or even be reinforced, creating a possible new high-stake assessment framework.

Session P - Formative Assessment III

9:00 - 9:30

Implementing grade-free schools: Justifications, challenges and opportunities from principals' perspectives

T. Burner[1], A. Gillespie[2]

[1]University of Southeast Norway, Norway
[2]Oslo Metropolitan University, Norway

This paper reports on four principals' views on the implementation of grade-free middle schools in Norway, i.e. schools that drop all grades on students' performances except the two required by the national Education Act, as part of implementing formative assessment practices. More specifically, we were interested in the under-researched area of how principals justify introducing and implementing grade-free schools, and what their experiences are regarding challenges and opportunities that have arisen during and as a result of the implementation. We chose the informants according to the following criteria: 1) the school had been giving grades at an earlier stage 2) the school is currently a grade-free school 3) the current principal had been responsible for the changes. Semi-structured in-depth interviews were used to collect data. Findings suggest that principals rely on research and unsatisfying assessment practices when justifying a change to grade-free schools. However, they do not find the involvement of students, nor the information directed at parents about the implementation, to be sufficient. They also mention challenges related to the current assessment system, which they believe underpin a behavioristic understanding of learning. We call for more research on trust among various stakeholders and student involvement when implementing grade-free schools.

The first stage of a CEFR standard setting procedure for four communicative skills of English in Kazakhstan

A. Dyussenova[1], D. Sartauova[1]

[1]Nazarbayev Intellectual schools, Kazakhstan

Nazarbayev Intellectual schools (NIS), a project initiated in 2008 by the First President of the Republic of Kazakhstan, support the development of three languages for different purposes: developing the state language, Kazakh, maintaining the development of Russian language as an additional tool for international communication, and English as a tool to gain access to opportunities at the international level. Learning these languages is not only promoted for communication purposes at NIS but also for use as a mean of instruction for content of other subjects such as Mathematics, Biology, Chemistry, Physics, Geography, National and World History, Politics and Economy. Studying other subjects in the three languages simultaneously helps to expand access to information, new perspectives, and a deeper understanding of other cultures. The trilingual policy contributes to the increase of students' capacities in critical and creative thinking as well as ability to intercultural cooperation.
The aim of the present paper is to present to the public eye the development of the monitoring system of second and foreign languages, a unique and an ambitious educational project in Kazakhstan, and to present the results of the first stage of a standard setting procedure in accordance with CEFR.

Challenges and opportunities for formative assessment practices of reading comprehension in vulnerable Chilean classrooms. A multiple case study

E. de Padua[1, 2]

[1]University of Cambridge, United Kingdom
[2]Agencia Nacional de Investigación y Desarrollo, Chile

This qualitative research project aims to analyse, through a multiple-case study, reading classroom assessment practices in Chile considering teachers' and students' perspectives. Additionally, it aims to identify specific challenges that need to be considered when designing support strategies for classroom assessment and teachers' professional development opportunities in this area.
This research considers five cases. The research methods included interviews, classroom observations and the analysis of assessment strategies. Each case was selected as an example of a formative approach to assessment and according to students' vulnerability indicators. Thematic analysis and discourse analysis were the tools used to analyse the data.
Among the main findings, there are: the contradiction teachers experiment when trying to motivate their students to read and -at the same time- assessing their reading skills, the omnipresence of grades as a factor that limits teachers' practice and students' learning, and the implementation of the formative approach predominantly from an instrumental perspective. Finally, socio-cultural elements of reading and assessment can be an opportunity for the implementation of formative assessment practices; however, the lack of trust in teachers forces them to keep a traditional approach in order to fit into the system.

Session Q - High Stakes Assessment

9:00 - 9:30

Teacher Assessment Identity in the Context of a High Stakes Examination

M. O'Leary[1, 2], Z. Lysaght[1], A. Doyle[1]

[1]Institute of Education, Dublin City University, Ireland
[2]Centre for Assessment Research, Policy and Practice in Education (CARPE), Ireland

The Leaving Certificate (LC) is a high stakes examination with results feeding into a point system that is used for entry to higher and further education. Irish society looks to the LC examinations as fair, reliable and the process carried out by the State Examinations Commission enjoys public support. However, in June 2020 Irish post-primary teachers and students, were faced with the reality that the high-stakes LC examination was being cancelled due to concerns around Covid-19. It was agreed by the education partners that the crisis would require that teachers estimate a mark and class rank for each of their students, followed by a process of national standardisation once data from all schools were submitted to the Department of Education and Skills. Known as calculated grades, the initiative was historic as, for the first time ever, Irish post-primary teachers were involved in assessing their own students for high-stakes certification purposes. Drawing on the work of Looney et al. (2020) and data from an online survey of over 700 post primary teachers, this paper explores how being involved in the calculated grades process impacted teachers' feelings and beliefs about their role as assessors.

Equal opportunity or unfair advantage? The use of test accommodations in high-stakes assessments

C. Vidal Rodeiro[1], S. Macinska[2]

[1]Cambridge Assessment, United Kingdom
[2]Cambridge Assessment English, United Kingdom

Using test accommodations (also known as access arrangements) as means of meeting the needs of all students and ensuring fairness is now common practice in many countries. The purpose of such accommodations is to provide all students with access to the assessments, enabling them to demonstrate their knowledge and skills by removing unnecessary barriers. However, there has been some controversy around the practice of providing test accommodations, with some critics suggesting that they may provide an unfair advantage, rather than simply levelling the playing field. If that were the case, the assessment results of the students with accommodations could be inflated, which could have a detrimental effect on the validity of the assessment. This research set out to investigate this claim using data from an international awarding body in the United Kingdom.
The performance of students who completed high-stakes examinations with and without test accommodations was compared. To account for group differences that have the potential to affect performance, students were matched on a number of background characteristics using a propensity score matching procedure.
The results revealed that students with and without accommodations performed similarly (received comparable grades in their assessments), suggesting that the test accommodations are working as intended.

Predictive validity in selection to higher education: potential barriers for students with immigrant backgrounds

M. Strömbäck Hjärne[1], C. Wikström[2]

[1]Umeå University, Sweden
[2]Umea University, Sweden

The ambition of many selection programmes is to broaden the student recruitment base and enhance inclusiveness. Students applying to higher education in Sweden are heterogeneous when it comes to language and cultural background. Still, the selection instruments for higher education used in Sweden are verbally demanding and require high proficiency in the Swedish language that potentially cause barriers for students who do not have Swedish backgrounds. In this study, comparisons were made between performance on the selection instruments and first year achievement in higher education for students with Swedish backgrounds and students with immigrant backgrounds. Data from 39 258 students admitted to 11 of the major programmes in higher education in Sweden was analysed using predominantly multiple regression analysis. Group differences in regression model residuals show that the selection instruments over-predict academic achievement in higher education for students with immigrant backgrounds.

Automatic marking of student essays: making sense of AI in assessment

11:00 - 11:30

Artificial Intelligence in the prime marking process – a comparison with current applications of machine learning

J. Burton[1]

[1]AQA, United Kingdom

Artificial Intelligence in the prime marking process – a comparison with current applications of machine learning
The process of prime marking is a ripe area for the application of artificial intelligence (AI). A machine learning algorithm has the potential to make the marking process more efficient, while at the same time providing a more uniform and objective measurement of candidates' responses.
In recent years, there has been a huge expansion in the application of natural language processing (NLP), together with sentiment analysis; several industries have taken a series of complex and interpretable information and extracted the key themes, issues and points from a large body of text.
We will present case studies from some of the industries that have applied AI to their core processes and consider how these could inform the assessment industry's approach to marking. Using examples from the tax and accounting sector and the NHS, we will focus on three key areas: translating and digitising long-form handwritten responses; understanding the sentiment of the person writing in long form; and assigning a value to the response.

‘Explainability’ of machine learning algorithms and implications for reviews of marking and appeals

C. Aloisi[1]

[1]AQA, United Kingdom

This presentation considers the challenges exam providers would encounter if post-results services for schools (e.g. review of marking or appeals) were carried out by an algorithm. We focus in particular on the issue of ‘explainability’: how decisions produced by an algorithm would have to be made understandable by humans. In England, current post-results regulations are underpinned by principles such as accountability, public trust and fairness; hence, if the use of machines was permitted in the marking review process, machine-to-human communication would be inevitable.

We look at how key concepts such as ‘interpretability’ and ‘transparency’ are currently understood and how they are implemented in practice. We also consider how effective widely-used methods such as LIME or saliency maps would be at dealing with the complexity of post-results services.

We move on to focus on the performance and capabilities of general natural language processing algorithms such as ELMo, BERT and OpenAI’s GPT-2. We consider how they could be adopted for post-results purposes. Using actual examinations and student responses, we illustrate the limitations of these algorithms with respect to explainability.

Finally, we note that current shortcomings of machine learning algorithms are often related to mark schemes and their interpretations. We therefore conclude that mark scheme research may regain prominence over the next few years.

Embedding an AI-based marking process into a high-stakes assessment

D. West[1]

[1]AQA, United Kingdom

This presentation considers how to integrate a language model that has been applied to a passage of free text into a framework for marking and ranking.
An AI system has to be trained for a marking task using a representative set of pre-marked responses. The responses must have been marked by at least two experts, with any disagreements resolved, and the marks screened for any bias.
To build a predictive system that can provide marks for previously unseen responses, the language model must be adjusted to guard against ‘over-fitting’ to the data. This adjustment is essential, and is achieved by tenfold cross-validation and an estimate of precision and recall. An automated second marker used for quality control might be optimised for recall, while a classifier will need more precision. A taxonomy of aspects of language modelling illustrates the main areas of focus needed.
An interpretable model, made possible through the use of prototype layers within a network, would enable methods such as k-clustering. Such a model could automate the selection of seeds for quality control. Any identified outlier responses – those that fool the initial AI marking model – could be exploited to build up resistance to such responses.

Exploring confidence in assessment practice

11:00 - 11:30

Trusting teachers' judgements: responses to COVID19 in England

M. Richardson[1]

[1]UCL Institute of Education, United Kingdom

England has witnessed a gradual erosion of trust in the teaching profession. This suspicion is notable when the results of the summer national examinations are released, but in 2020 the pandemic hit our awarding systems. Suddenly, teachers' professional evidence became valuable currency for students due to undertake post-16 study, employment or university. Then dramatic shifts in policy revealed a government that was under confident and fearful of a public outcry about examination results. They were right to be afraid because the narrative that emerged through the press and social media was based on themes of inaccuracy, mutating algorithms and a challenge to define what was fair about their decision-making.
Particular narratives relating to trust, fear and bias continue to emerge from the debates surrounding the policy actions that led to the cancellation of examinations not only in England, but worldwide. This paper considers the unpredictable outcomes of the reforms to assessment and the particular discourses that characterised attitudes of confidence in teacher assessment in high stakes environments. COVID presents an opportunity to change how we view education and most importantly, to understand what invokes confidence in assessment practices so they best support the aims of education systems.

Rise of the machines? The evolving role of Technology and Artificial Intelligence (AI) technologies in low and high stakes assessment in the UK

R. Clesham[1]

[1]Pearson UK (corporate membership), United Kingdom

This paper considers how research and advances in technology and Artificial Intelligence (AI) are viewed and understood in the context of educational assessment in the UK. Despite a prevalence of AI and related technologies in everyday life, within the context of low and high stakes assessments in the UK, the use of technology (particularly those including AI) has evolved slowly. This is surprising because there is obvious value in harnessing contemporary computing power to facilitate AI and automated machine decision-making in the marking and processing of assessment data and these systems are already successfully implemented in globally used language tests. We might expect that the functionality of AI technologies alone would be a verifiable means to improve assessment workload and logistics. However, there are several issues at play here, particularly concerns about the efficacy of reliance on automated systems to facilitate authentic assessments of student learning in formative and summative settings. There are also pragmatic issues; the fear of machines taking over human roles, which could potentially leave teachers and assessors redundant. The discussion considers the nature of trust in the use of AI and technology in assessment and asks: How do we make it work effectively for our purposes?

What builds public confidence in the fairness of standards in high stakes assessment

J. Baird[1], L. Hayward[2], M. Ware[3]

[1]University of Oxford, United Kingdom
[2]University of Glasgow, United Kingdom
[3]Scottish Qualifications Authority, United Kingdom

Scotland, traditionally, has high levels of confidence in teachers. Fairness and Justice are key concepts in how Scottish Education is presented, as reflected in the myth of the ‘Lad o’ Pairts’, the child of humble origins succeeds in life because of educational opportunity. For more than 100 years, the high stakes assessment system in Scotland, with the Scottish Higher at its heart, has been crucial to that sense of opportunity and justice. However, in session 2019-20 and in common with concerns in other parts of the UK, public confidence in the High Stakes Assessment in Scotland was dented; nationally moderated results were rescinded and replaced by results based on locally moderated teachers’ professional judgment. This paper reports on a participative research project involving the Universities of Glasgow and Oxford and the Scottish Qualifications Authority that sought to understand public perceptions of standards and fairness across a range of key communities, from students to employers. Reporting on evidence emerging from both qualitative and quantitative data, the researchers reflect on key insights emerging and on how participative research models might be used to improve the quality of understanding across policy and practice communities in highly charged educational environments.

Impact of the disruption to schooling due to the pandemic - data from England

11:00 - 11:30

A study of the impact of the pandemic on the attainment of students aged 5–7 in reading and mathematics

S. Rose[1]

[1]National Foundation for Educational Research, United Kingdom

This paper will present how the learning of students aged 5-7 may have been impacted by school closures due to Covid-19. Standardised assessments in reading and mathematics measured the performance of 12,000 students. The research builds a holistic picture of students' school experience in 2020/21 including the development of social skills and the effectiveness of the intervention strategies intended to help them recover. The paper will present data from the assessments taken at three time points in the year. It will present the quantitative evidence of the impact of the disruption on attainment overall and between subgroups of students. Interim findings published throughout the year gave schools timely assessment and diagnostic information and contributed to the fast growing evidence base of how students have been affected.
Whilst it was possible to quantify the impact of the school closures on students' learning, knowing the gap is only part of the way in which assessment can inform teachers. The paper will introduce the rationale for including diagnostic assessment to create a detailed picture of the impact on skills development.

Diagnostic assessment information for teachers – reading

T. Paxman[1]

[1]National Foundation for Educational Research, United Kingdom

This paper will go on to expand on the headline findings from the main study through the lens of the diagnostic analysis to show the more specific ways in which different groups of students were affected in reading. For example, in wave 1 (autumn) it was observed that 6-7-year-old students had made, on average, two months less progress over the last 12 months than would be expected. Diagnostic analysis showed which skills students found difficult in 2020 and highlighted patterns of performance for specific groups of students. Further scrutiny revealed a differential impact with the greatest effect on students in the early stages of learning to read i.e. with limited ability to read independently.
This paper will also show how the development behind the standardised reading assessments, the range of item types and curriculum areas included, make diagnostic conclusions about students' performance possible. In this session, we also hope to showcase the use of this working document throughout the academic year 2020/21, highlighting the ways in which different assessment tools can support the education community. The audience will be encouraged to explore possible reasons for these findings and discuss the implications.

Diagnostic assessment information for teachers – mathematics

P. Akhtar[1]

[1]National Foundation for Educational Research, United Kingdom

The outcomes of the mathematics diagnostic assessment will be used to frame the headline findings from the main study. In particular, students' performance across different mathematics topics will be compared to reveal how, despite the 2020 cohort being around 2 months behind those pupils in the standardisation sample, they actually performed better in some areas.
This session will also show how diagnostic assessment and the frames used to code responses were able to delineate performance in even more detail, showing specific areas of different mathematics topics which students found more difficult. Furthermore, the coding of responses investigated students' use of different strategies, such as marks to enable counting or drawing a number line. This offers interesting insights into how students approached different items, and how this may have changed as students progress through the academic year. The audience will be invited to debate why students' performance may have varied across mathematics topics, particularly considering the context of school closures and remote learning.

Session R - Language Issues in Assessment

14:00 - 14:30

The Item-Explanatory Models with Language Features for a Test of Russian as a Foreign Language

M. Skryabin[1], M. Lebedeva[2]

[1]Independent Researcher, Russia
[2]Pushkin State Russian Language Institute, Russia

The language placement test, intended to diagnose the test-taker level of language competence, is a specific type of language testing. The accuracy of measurement is determined by the quality of the instrument itself, and by the psychometric properties of its items. In this work, we consider how the linguistic features of the items relate to their difficulty on the example of a grammar-and-vocabulary test on Russian as a foreign language. High-level text features, such as narrativity, coherence, etc., in this case, are not to be considered, whereas low-level characteristics such as lexical features, grammar topics may be a factor determining the item difficulties. The test contains 120 questions, marked according to the four proficiency levels of RFL from A1 to B2. The questions are aimed at revealing the acquisition of grammar topics and vocabulary items, which are studied at each level. For analysis, we used the linear logistic test model with item error (LLTM+e) that considers the uncertainty in explanation and enhances the precision of estimation of the item difficulties by accounting for residual variation. Our results show that the effect of two topics, Verb (vocabulary) and Sentences, were significant when the Case topic was set as a reference level.

Different but the same: Score concordance across academic English language proficiency tests

S. Hughes[1], R. Clesham[1]

[1]Pearson, United Kingdom

In the context of international high stakes English language tests, scores from different tests are often used for the same purpose. For example, university admissions or immigration visa applications may require applicants to demonstrate a specified level of English language proficiency. These institutions may recognise a number of different English language tests as appropriate means to demonstrate proficiency, though there is no central regulator that ensures comparability between these assessments. In this case, assessment organisations must conduct alignment research to support the interpretation and use of test scores. Linking studies can be used to produce a score concordance table, which identifies comparable scores on similar tests produced by different testing organisations.

This presentation focuses on the results, challenges, and implications of a linking study between PTE Academic and IELTS Academic. Scores for these tests carry significant currency in terms of academic, professional and economic migration entry requirements and a score concordance table is essential to support test score users in making valid interpretations of test scores and fair decisions regarding test takers.

Session S - Psychometrics II

14:00 - 14:30

Catching the change of test difficulty with multiple methods: equating Cyprus teacher admission examinations

M. Van Onna[1], I. Lamprianou[2], C. Jongkamp[1]

[1]Cito, Netherlands
[2]University of Cyprus, Cyprus

In 2019, the second wave of placement examinations for the Cyprus teachers' appointment system took place. Candidates received a final Uniform Mark Score (UMS). The UMS of 2017 and 2019 form one ranking of all candidates, from which teachers are selected if positions are available.
Several quantitative and qualitative methods have been used to assess the amount of difference in difficulty of 32 subjects. The quantitative methods that were applied, comprised resit analysis, propensity score matching, the equivalent groups method and the pseudo anchor method. The computations of these methods use the scores of the candidates. The qualitative methods that were used, were the statistically informed standard setting, the explicit expert judgment and the implicit expert judgement. These qualitative methods are based on expert judgements.
Typically, two to four methods were applied per subject. The estimates of the applied methods were averaged to form a combined estimate of the difference. The inverse of the standard error of each method was used as the weight in the averaging.
Details about the precision of equating and other technical issues about equating will be discussed at length. We conclude with practical recommendations both for researchers and practitioners.

Flexible and secure high stakes testing using Linear-on-the-fly-Testing

S. de Klerk[1], A. Verschoor[2]

[1]eX:plain, Netherlands
[2]Cito, Netherlands

The main strategy for society to cope with the Covid pandemic has been a rapid digitization. Examination is no exception to this phenomenon. Flexibilization and security may well become even more important in the near future. Linear-on-the-fly-testing (LOFT) is an option with advantages, especially for examination: Candidates will all take a unique exam form without the technical complications of computerized adaptive testing (CAT). A further advantage of LOFT over CAT is, that it is possible to work with a single fixed cut off score instead of hard-to-explain ability estimates.
The easiest, and most popular, implementation of LOFT is a random selection of items according to the test blueprint. Yet, the question remains whether random selection is fair for the candidates as the difficulty level may vary. Automated Test Assembly may be a better option.
In this presentation, we compare two automated test assembly (ATA) models with these random selections. Not only the difficulty of the exam forms will be controlled to about any given margin, also item exposure can easily be controlled, thus making it possible to find an optimal compromise between exposure control and reliability.

Session T - Covid-19 Research III

14:00 - 14:30

Setting and maintaining standards in technical qualifications in England before and after Covid-19: evidence and some reflections

N. Zanini[1], T. Stratton[1]

[1]Ofqual, United Kingdom

Between 2017 and 2020 Ofqual, the Regulator of Qualifications in England, conducted a programme of work to inform the development of an awarding methodology for the maintenance of standards in Applied General and Tech Level qualifications. The aim of this presentation is to report the range of quantitative analyses conducted to this purpose and to evaluate the introduction of the methodology to set and maintain standards on a small scale. The empirical work was based on a rich dataset on students' performance in these qualifications linked to a broad set of students' characteristics, including prior schooling and socio-economic background. The findings will be used to reflect on the introduction of an awarding methodology based on prior attainment-based predictions for technical qualifications. Given that in summer 2020 the cancellation of exams in England stopped the plans to roll out the methodology on a larger scale, the empirical evidence will be also discussed in relation to the impact of Covid-19 on the use of assessments and awarding methodologies that may result in increased fairness to learners.

A year of pandemic: managing the impact in Power Maths primary schools in England, 2020-2021

E. Barrow[1], J. Golding[2], G. Grima[3]

[1]Pearson Education, United Kingdom
[2]University College London Institute of Education, United Kingdom
[3]Pearson UK, United Kingdom

Research on the impact of the Covid-19 pandemic on primary-aged children's learning paints a discouraging picture. Studies have highlighted mathematics as one of the most heavily affected areas. We report on findings from a longitudinal study of use and impact of Power Maths, a 'mastery'-oriented primary (R-year 6) mathematics resource, in England. The study follows 40 classes of 2019-21 Power Maths-using (initially) year 1,3 and 5 children and their teachers over two years, exploring teacher and pupil use and impact of that on mathematics learning. We discuss the response of teachers to the changing demands of teaching, learning and assessment throughout the pandemic, from Summer 2020 to Summer 2021 and through four cycles of data collection. Initial responses from teachers during Summer 2020, following mandated school closures, identified challenges in reliably assessing primary pupil progress in mathematics learning during the period of home-learning. Study data from Autumn 2020 showed varied approaches to assessment of pupils once schools had reopened in September, with adaptations made in the classroom both practically and pedagogically, and then further adaptations during the second sustained lockdown period January-March 2021 and the ensuing still-constrained period. Challenges in reliably assessing younger children's pandemic-constrained work have reduced, but persist.

What is lost when exams are cancelled?

B. Redmond[1], J. Golding[2], G. Grima[3]

[1]Pearson, United Kingdom
[2]University College London Institute of Education, United Kingdom
[3]Pearson UK, United Kingdom

We use data from England, where high-stakes GCSE (age 16) and A level (age 18) examinations in summer 2020 were cancelled due to Coronavirus, to highlight direct impacts on mathematics students' learning. The source longitudinal (2017-21) classroom-close study focused on students' and teachers' experiences of newly reformed mathematics A levels, but was adapted in 2020 to explore the impact of Coronavirus. In 2020 we also collected data from academics to capture students' preparedness for progression to higher education.
We show that preparation for examinations supports students to develop, consolidate and synthesise learning. Where examinations were cancelled most mathematics GCSE and A Level students were unable to take advantage of this intensively fruitful period of study. We also evidence missed, and lamented, opportunities for students to gain experience of revision techniques and exam practices.
Finally, we discuss the role of examinations as an important ritual, allowing students to feel their learning has been formally and objectively validated. Linked to this, we saw some students' confidence reduced and an increase in students experiencing 'imposter syndrome'. Findings have implications for students' progression - from GCSE to A level and from A level to university - but also for the future of assessment.

Session U - Fairness & Social Fairness III

14:00 - 14:30

The impact of representative subject content on the attainment gap: Ethnicity and Outcomes in GCSE History

K. Mason[1], E. Barrow[2], M. Hill[2]

[1]Pearson, United Kingdom
[2]Pearson Education, United Kingdom

One of the recommendations of the MacPherson report was to consider how the National Curriculum in England could better reflect the needs of a diverse society. However, twenty years later, the ways in which GCSE History curricula respond to a culturally diverse school cohort is still being debated and challenged. Education is crucial to ensuring greater social equity amongst people of different backgrounds. However, research has shown that the attainment gap between students of different ethnic backgrounds has been persistent over the last ten years.
This paper uses the UK's National Pupil Database to look at the attainment gaps between students of different ethnic backgrounds as seen in Pearson's GCSE History assessments in Summer 2019, and links these to the subject content of the options chosen in order to explore whether a representative curriculum is one avenue for addressing these and creating more equitable outcomes.
This study is the first of an intended program examining the impact diverse subject content has on student outcomes. GCSE History has been selected because of public interest in race and history teaching and because of the structure of the qualification.

Does removing tiering from high-stakes examinations reduce the size of attainment gaps?

M. Carroll[1]

[1]Cambridge Assessment, United Kingdom

In England, Wales and Northern Ireland, the General Certificate of Secondary Education (GCSE) was introduced in the 1980s to replace separate qualifications for higher- and lower-attaining students. To allow all students to access GCSEs, many subjects had tiered examinations: lower-attaining students took foundation papers, targeting lower grades, and higher-attaining students took higher papers, targeting higher grades. However, tiering was criticised for limiting attainment and aspiration of foundation tier students. Moreover, some argued that “capping” disproportionately affected certain demographic groups, thus exacerbating attainment gaps. Accordingly, in reforms in the 2010s, tiering was removed from most subjects. This provides an opportunity to examine the effect of tiering on attainment gaps. Here, attainment gaps between age, sex and socioeconomic deprivation groups were calculated for subjects where tiering was removed or retained, or which were never tiered, for years before and after reform. Against expectations, removal of tiering was not associated with reduced attainment gaps; indeed, gaps often increased, in contrast to subjects that retained tiering. These effects cannot be definitively attributed to tiering, but results indicate that impacts of tiering removal may be more complex than anticipated, and further research may be required to understand responses of schools and students to untiered examinations.

Sesion V - Validity & Validation

14:00 - 14:30

Comparability of computerized performance-based assessment for measuring critical thinking

D. Gracheva[1]

[1]Higher School of Economics, Russia

A broad use of performance-based tasks (PBTs) in educational assessment lead to a necessity to construct different forms of the same task. However, the methodology of comparability of PBTs is underdeveloped and there is a shortage of research answers this issue. The purpose of the paper is to provide comparability evidence of performance-based assessment. In this study, we used test data from computerized PBT aimed at measuring critical thinking of primary school children (n = 537). The comparability of PBT forms were examined using within-subject design. The PBTs were developed following Evidence-Centered Design. In the original form students are asked to find information about aquarium design for crabs and then equip the aquarium. In the second form students do same steps to make a terrarium for geckos.In order to provide comparability evidence of PBT forms methodology of invariance testing with CFA was used. Configural and strong invariance was tested across task forms along with the mean comparison of latent factors. Findings indicate that PBT forms can be considered quite comparable on main factors related to critical thinking. However, context factors appeared to be easier in the “Terrarium” task, demonstrating the influence of context on the comparability of PBTs.

Validation of the updated student selection system to enter grade 7 of Nazarbayev Intellectual schools: effectiveness and predictive validity

A. Issatayeva[1], A. Jandarova[1], Z. Jumabayeva[1], Y. Nurguzhin[1], N. Dieteren[2], F. Kamphuis[2]

[1]Nazarbayev Intellectual schools, Kazakhstan
[2]Cito, Netherlands

Nazarbayev Intellectual schools (NIS) are a network of STEM schools that selects talented and motivated students.
This paper introduces the 2020 validation research for the system of development and implementation of a testing system for selecting potential students to study at NIS. For the first time in 2015, a validation research was conducted to justify that the developed assessment systems are fit for purpose and the right students were selected to NIS during the selection procedure. Since then, NIS student selection system has undergone changes.
In the paper we will particularly focus on the longitudinal research of the effectiveness and predictive validity of the student selection system for 2,925 students in grade 7 from 19 schools. The data for the validation research were based on the assessment results from one cohort of NIS students in the academic year 2019-2020.
The paper concludes that even though the student selection test changed starting from 2019, this did not influence the power of predictive validity of the subtests in general.

Development and Validation of Competency Frameworks: aligning purposes, constructs, structures and methods

S. Child[1], S. Shaw[2]

[1]Cambridge Assessment, United Kingdom
[2]University of Cambridge, United Kingdom

The concept of ‘competence’ has gained traction in recent years because of its power in informing educational objectives, such as readiness for transition to later educational stages. This paper aims to provide a theoretical and practical steer for competency framework developers, by arguing for a clear alignment between competency framework purpose, competence definition, framework structure, and validation methods.

We suggest three key distinctions that developers can make in terms of the structure of a competency framework: binary vs continuum; atomistic vs holistic; and, context-specific vs context-general. Decisions in early stages of framework development inform the range of framework claims. Analysis of validation case studies revealed three main thematic issues: instrument-focused validation; group interactions biasing validation outcomes; and, limitations in terms of validation evidence collection from framework users. We argue that to resolve these issues, developers need to understand how the claims made in relation to a competency framework link to the scope of validation enquiry, and the on-going engagement of framework users. We conclude with a description of a template of questions for developers to consider in determining the range of potential claims to be made concerning their framework, and understanding competency framework users and contexts.

Session W - Assessment Cultures I

14:00 - 14:30

Defining and operationalising synoptic assessment

F. Constantinou[1]

[1]Cambridge Assessment, University of Cambridge, United Kingdom

Synoptic assessment refers to the assessment of students' understanding of the connections between the different elements of a subject. Despite constituting a useful pedagogical tool (e.g. it can encourage a holistic and coherent delivery of the curriculum), synoptic assessment remains an under-researched topic in the field of educational assessment. In fact, an electronic search for the term synoptic assessment in major academic databases yielded only four related resources. To address this gap in research, this study examined synoptic assessment both at a theoretical and at a practical level. First, it attempted to illuminate the concept of synoptic assessment by situating it in the broader educational literature. Subsequently, it sought to operationalise it by developing a framework that captures the different routes via which synoptic assessment can be incorporated into the design of various types of courses. This paper will unpack the concept of synoptic assessment and will describe how synoptic assessment can be introduced into academic and vocational courses.

Assessment Literacy Enhancement of language teachers: How can we support them?

T. Rousoulioti[1], D. Tsagari[2]

[1]Aristotle University of Thessaloniki, Greece
[2]Oslo Metropolitan University, Norway

Teacher assessment is an important component of the educational process. In Greece, in particular, language assessment is enacted in a multitude of ways, e.g. prepare students for large-scale exams; design and score a wide variety of language tests; provide useful feedback to learners based on results of such assessments; align their assessment procedures with language curricula to meet national or European language assessment standards. However, research has shown that teachers find it challenging to fully address their assessment mandates.
The aim of this study is a) to investigate the level of language assessment knowledge and skills of pre- and in-service teachers of Greek as L2 and b) train them in assessment aspects. Results showed that teachers followed a summative orientation to their assessment with strong preference for online training in assessment while the training material used during the intervention phase managed to support the majority of the participants and to meet their assessment needs. Additional findings relate to the effectiveness and sustainability of the trainijng materials in educating teachers in language assessment. The current paper informs and expands very well-known conceptualisations of LAL and makes research and pedagogical recommendations in enhancing teachers' language assessment literacy.

Assessment Literacy – How does being an examiner enhance teachers' understanding of assessment?

V. Coleman[1], M. Johnson[1]

[1]Cambridge Assessment, United Kingdom

Concerns have been raised that many teachers do not have sufficient Assessment Literacy (AL) which has implications for teacher professionalism and practice. AL encompasses the basic understandings, skills, and applications that underpin a teacher's ability to use and understand assessment, as well as teacher's beliefs and feelings about assessment. This means that the relationship between AL and assessment practice is complex and multidirectional. Teacher AL is more important than ever in these changing times, with teacher assessment crucial for an effective education system.

When investigating teacher AL, the metaphor of an 'assessment career' is useful. AL is changeable over time and is influenced by both personal and professional experience. Teachers' participation in formal examining is one experience that may influence their AL and their wider practice.

To explore the influence of examining on teacher AL we used concept maps and interviews with a sample of Science and English teacher-examiners. Using this we developed a survey to explore the influence of examining on the development of AL amongst a wider sample of international teacher-examiners. The outcomes of our study discuss the contribution that professional examining work has on transforming teacher's AL and the impact on their teaching practices.

Ignite Session

16:00 - 16:10

“Assessment for Changing Times: Opportunities and Challenges” – The AEA-Europe eAssessment Special Interest Group (SIG)

M. Ware[1]

[1]Scottish Qualifications Authority, United Kingdom

This presentation will explain how the AEA-Europe eAssessment SIG, established in 2017, supports sharing experiences in the effective use of e-assessment. It will outline work to date and future plans and encourage all members to help shape and contribute to SIG activities.

Fostering the student-teacher relationship in the age of distance learning: the use of screencasting to provide assessment feedback

S. Nuseibeh[1]

[1]University College London, United Kingdom

In Higher Education, sadly, we often hear students complaining about the quality of feedback they receive on assignments with many suggesting that the traditional, written feedback style can come across as meaningless and impersonal.
There is an increasing amount of data in the literature to support the idea of moving away from written feedback in favour of providing more informative summaries through the medium of video (which is capable of providing both auditory and visual cues to the student). Video-based feedback can come in a few different forms, but all have been shown to strengthen the student-marker relationship with a high level of acceptance for both students and staff.
This presentation will demonstrate how a “combination screencast” technique can be used to provide personalised and meaningful feedback to students on their assignments and how it can be used to foster the student-teacher relationship in the age of distance learning.

Digital Storytelling as a Technology-enhanced Assessment Technique

S. Caspari[1]

[1]Uni Passau, Germany

Digital Storytelling (DST) has gained attention in higher education due to its potential to capture lived experience, engage students with content, promote reflection and creativity. Although available literature is abundant with the use of DST as a teaching and learning strategy, there are few studies on its usage as an assessment technique.
This study aims to introduce using DST for different assessment purposes, namely formative, summative and integrative. First, DST is presented as a formative technique which simultaneously promotes and assess Project-based learning. Second, DST is discussed as a summative assessment technique to measure Multimedia Literacy skills. Eventually, DST is explored as an integrative assessment technique (assessment-as-learning) by capturing students' metacognitive skill (reflection-on-learning) in e-learning courses. For each usage, Best Practices as well as potential threats to assessment qualities and suggestions for improvement will be characterized.

An exploration of the effects of examination speededness in A level mathematics, A level science, GCSE mathematics and GCSE French

Q. He[1], B. Black[2]

[1]Office of Qualifications and Examinations Regulation, United Kingdom
[2]Ofqual, United Kingdom

Test speededness refers to situations where a substantial proportion of the examinee population cannot complete the test within the specified time limits and can be detrimental to its intended functioning. In this study, the effects of speededness in fifteen secondary school examinations from England were explored using two approaches, one based on the relationship between scores on speeded and non-speeded items and the other based on mixture Rasch modelling. There appears to be a certain amount of speededness in the exams investigated, with two of the mathematics exams showing the largest effects. Maximum item omission rates vary from 11% to 57% in the exams. Percentages of total number of items that were classified as speeded items range from 11% to 50%. Percentages of maximum exam marks associated with speeded items vary from 7% to 65%. It was also found that, other things being equal, candidates with lower abilities generally had higher probabilities of being classified into the most speeded groups than candidates with higher abilities. The effects from other variables such as gender, status of special educational needs, eligibility for free school meals, ethnicity, and home first language were found to be insignificant for most of the exams studied.

Supporting schools in the analysis and use of external assessment reports

A. Lobo[1], M. Borges[1], A. Monteiro[1]

[1]Institute for Educational Assessment, Portugal

With the PAR project, IAVE aims, in partnership with schools, to analyse and reflect upon the best way of using the external assessment reports to provide added value to students' learning. The focus is standardised test results reports (low stakes external assessment tests) of various subjects and curricular areas taken by 2nd, 5th and 8th grade students (individual standardised tests reports and standardised tests school reports).
The first phase of the project (2019-2020) involved 17 schools in Portugal and a Portuguese school in Cape Verde. The sessions with head teachers, teachers, parents and students directly involved in these tests saw participants sharing experiences on using the information in the reports, and identifying the difficulties arising from their analysis, thus helping with pedagogical application.
The second phase consisted of 18 short training courses for teachers working in Portuguese schools, to disseminate IAVE's main recommendations to analyse and use these test results.
In 2021, the project is expected to be extended to more schools and a report will be published as a manual of good practices and guidelines, so that schools can operationalise and use the results generated by the external assessment tests effectively and in a pedagogically relevant manner.

A teacher-report measure of student's personal social-emotional skills: a multilevel mixture IRT approach

I. Uglanova[1], A. Ivanova[1]

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

Social-emotional skills (SES) are crucial for children’s positive behavior in school. The main aim of this study is to investigate the patterns of children’s social and emotional development and describe distinctive groups of children presented as latent classes who have similar non-cognitive characteristics. Identifying specific patterns in children’s development may help teachers to create strategies that will address the unique needs of children. To reach our aim we used a sample of nearly 5000 first-grade students recruited from 195 schools in one of the regions of Russia. We applied a multilevel mixture model characterized by an IRT parameterization, specifically, multilevel LC IRT, which allowed us to take into account the hierarchy of our data. We obtained five homogeneous classes at the lower-level (students) and two homogeneous classes at the higher-level (teachers). We found that gender, math, and reading scores are significant predictors of the latent classes at the lower level. The higher-level latent classes significantly differed by teachers' self-efficacy attitudes, whereas other covariates were not significant predictors. It is also shown that the communication subscale items demonstrate different difficulties for latent classes.

Is the inclusivity paradox of technology in education holding back onscreen assessment?

H. White[1], M. Campbell[1]

[1]Pearson UK, United Kingdom

Across primary, secondary and higher education, the education community was forced to seek new solutions to teaching, learning and assessment given the challenges presented by COVID-19. With that in mind, how do we seize the opportunity to advance the innovation and implementation in onscreen assessment?

The traditional context of high-stakes assessment in England dictates that in almost every instance examinations are paper-based. We are eager to explore the potential for maximising technology in assessment, whilst being mindful of the variation in provision across the education experience, both within and outside of school. It is, therefore, necessary to be cognisant of practical considerations needed to ensure comparability across national assessments and with paper assessments, as well as a viable solution for students who do not have digital access outside school.

Can we build accessible onscreen assessments that can be adopted successfully by schools? And, in doing so, explore new opportunities that enable a real paradigm shift with much broader benefit for teachers, learners and parents?

Session X - Assessment Cultures II

9:00 - 9:30

Assessment for changing times: opportunities and challenges

E. Andressen[1]

[1]Andressen Byram Ltd, United Kingdom

A key driver of shorter cycles of change and reform in assessment practice is the increased capability and use of technology for learning and assessment. In late 2020, Ofqual published a report into existing barriers to adoption of online and onscreen assessment, and how these might be overcome. During the pandemic, with high-stakes examinations interrupted and cancelled, calls for paper-based examinations to be replaced by an on screen alternative increased significantly. This paper will explore the extent to which these barriers have been overcome, and what remains to be done to achieve nationwide adoption of on screen delivery of high-stakes assessments in schools. It will review the barriers through the lens of 11+ secondary schools' entrance examinations in England. These were revised in content and delivery method at short notice when on site paper-based examinations became impossible. Their decentralised nature, and teachers' varied knowledge and experience of assessment, resulted in a range of alternatives and impacts on learners. There is some way to go but change is happening very quickly. It may not be possible to keep up the rate of momentum created by the pandemic but there is an opportunity to learn from 2020 to accelerate future adoption.

‘Elasticity’ in the administration of national assessment systems

A. Watts[1]

[1]University of Cambridge, UK, United Kingdom

At the end of the nineteenth century the government body which recommended the setting up of a school examination system in England and Wales proposed that it should align with “the English conception of variety and elasticity in educational organisation” (Bryce Report, 1895). This appeal to the national cultural context was a recognition that an examination system will be a reflection of the society which brings it into being. In the nineteenth century there was in Britain a deep-rooted resistance to government ‘interference’ in education. Support for the independence of secondary schools and teachers was a force which the government could not ignore and comparisons were made with the education systems in France and Germany which were described as too ‘centrally-controlled’. The word ‘elasticity’ therefore can be associated with an intention to decentralise control of the examination system.
This paper will review the implications of the word ‘elasticity’ at the end of the nineteenth century and will seek to explain what it came to mean in practice. It will then ask what the concept of elasticity has to say to us today about the challenge of centralisation, particularly when assessment systems are faced with introducing new ideas, practices and methods.

Session Y - Educational Policy and Assessment

9:00 - 9:30

Lesson observations for better teaching: Evidence from school inspections in Nazarbayev intellectual schools

G. Zhailauova[1], R. Kakabayeva[1], G. Kurmanbayeva[1], O. Mozhayeva[1]

[1]Autonomous Educational Organisation Nazarbayev Intellectual Schools, Kazakhstan

Investigating effects of lesson observations on teaching practices has occupied a prominent role within the domain of Education. Although the lesson observation has its main objective apparent in performance evaluation, it is a widely held view that feedback on observed lessons is among the most important factors for transforming teaching.
This study draws on two years' experience of lesson observations conducted within the school inspections in ten Intellectual schools in the Republic of Kazakhstan. The study aims to develop an understanding about the nature of lesson observation schemes developed within the school inspection framework and examines the role of feedback in improvement of teaching. Specifically, attention will be directed towards examining the role of lesson observations conducted by external evaluators, with discussion directed towards how feedback on teaching is connected to the development of teaching practices.
The classroom observation was used as primary method to evaluate teaching practice and generated valuable data on practical implications for teaching development. Additionally, a questionnaire-based survey was employed to establish association between lesson observations and improvement of teaching. The findings of this study complement those of earlier studies on this subject and suggest strong association between feedback on teaching and improvement of teaching practices.

The potential impact of unconditional University offers on A level attainment in England: evidence to inform the debate on proposed changes to University admissions

R. Taylor[1], N. Zanini[1], M. Walter[1]

[1]Ofqual, United Kingdom

Most 18-year-olds in England currently apply to University ahead of sitting their A-levels. Applications are based on grades predicted by schools, and Universities make an offer of a place to individual students. Such offers are generally ‘conditional’ on students achieving certain grades in their A-levels. In recent years, however, there has been a rise in ‘unconditional’ offers – offers that do not depend on grades and essentially guarantee students a University place.

This rise has raised concerns that students with unconditional offers will be less motivated to study for their A levels and might therefore under-perform. This could have implications for students’ attainment at A level and beyond.

To explore this, our analyses consider the possible implications of unconditional offers for students’ A level attainment. Our analyses use regression techniques to control for background variables that might be related to performance (prior attainment, gender, ethnicity etc.), and we analyse data across multiple years to consider any changes over time as the number of unconditional offers has risen.

The discussion focuses on the possible implications for student attainment, and our findings are considered in light of proposed changes to University admissions in England to a system of post-qualifications admissions.

How good can marking be? Adventures in marking, chapter 6 (or thereabouts)...

S. Holmes[1], B. Black[1]

[1]Ofqual, United Kingdom

The marking literature has established that marking consistency is lower for subjects with extended response items and divergent outcome spaces, with the probability of receiving the ‘definitive grade’ in live marker monitoring data varying from .96 for mathematics down to 0.52 for some English language and literature qualifications. At the same time, the desired precision for grading high stakes examinations is the same for all subjects given that their outcomes are used in the same way. This study tries to find out the extent to which marking consistency in a more ‘subjective’ subject could be improved. Just how good can marking be?
We looked to improve marking consistency on a GCSE English language paper using three stages of intervention. First, mark schemes were redesigned on the basis of research literature and expert input. Second, marker recruitment was highly selective. Finally, markers were given extensive, in-depth and interactive face-to-face training over a five day meeting.
Marking consistency improved by around 5-15% for most of the questions, and further analysis suggests that this is a conservative estimate. Therefore, whilst live marking consistency is good, there is some room for improvement, although there may be resource implications.

Session Z - National Tests and Examinations

9:00 - 9:30

Standardised Testing in English Reading and Mathematics in Irish Primary Schools: Trends Over Time

Z. Lysaght[1, 2], G. Cherry[3]

[1]Dublin City University, Ireland
[2]Centre for Assessment Research, Policy and Practice in Education (CARPE) at DCU, Ireland
[3]Centre for Assessment Research, Policy and Practice (CARPE) at DCU, Ireland

This paper presents findings from secondary data analysis of two large-scale surveys, conducted in Ireland approximately 50 years apart, that examined primary teachers' attitudes to and use of standardised tests to inform teaching and learning. Drawing on secondary data analyses of the quantitative data from both studies, and in particular comparison of responses to common survey items, this paper explores a number of overlapping themes relating to (1) teachers' confidence in test results, (2) practices and attitudes regarding the sharing of test data with stakeholders, (3) test preparation practices and (4) the influence of test data on teaching and learning. Based on a critique of the policy contexts that prevailed in Ireland when the respective studies were undertaken, the paper offers insights into how assessment policy has evolved to the point where standardised testing is gaining increasing prominence despite acknowledged commitment from all key stakeholders of the need for balance and alignment between assessment, learning and teaching.

Exploring Primary School Teachers' Use of Assessment Data in an Irish Context

P. Lehane[1, 2], V. Pitsia[1, 3], A. Karakolidis[4]

[1]Dublin City University, Ireland
[2]Centre for Assessment Research, Policy and Practice in Education (CARPE), Ireland
[3]Centre for Assessment Research, Policy and Practice in education (CARPE), Ireland
[4]Educational Research Centre (ERC), Ireland

Evidence suggests that the quality of a teacher's instructional practices can be improved if they are informed by relevant assessment data. The aim of this research study was to examine what factors can predict teachers' use of assessment data for instructional purposes through a secondary analysis of survey data from Ireland. This cross-sectional survey aimed to gather information about Irish primary teachers' use of, experiences with and attitudes towards one particular form of assessment data, standardised tests, in relation to reading and mathematics in Ireland. The current study sought to identify if teachers' reported use of assessment data for instructional purposes was related to their previous engagement with professional development on and attitudes towards assessment data use, taking into consideration relevant teacher and school characteristics. The analysis revealed that a positive attitude towards standardised tests as a form of assessment data and previous engagement in some form of professional development on the use of standardised testing during their careers were statistically significant predictors of assessment data use, explaining 11% of the variance in teachers' use of assessment data. This has implications for policy and practice which will be discussed.

National contest of thinking skills – an initiative to address the importance of 21 century skills development. Lithuanian experience

E. Melnikė[1], D. Sevalneva[1]

[1]National Agency for Education, Lithuania

Thinking skills, problem solving, critical thinking, creativity, are among 21 century skills, which are essential to develop in order to help our young generation to become ready for their future, lifelong learning. In order to address this challenge and to draw the attention of teachers, students, parents, other education professionals to the importance of the development of such skills, a national initiative The National Contest of Thinking Skills (NTC) started in Lithuania.
Since 2014, NTC in Lithuania is organized once every year. The main features are: participation is voluntary, only e-tests, the contest covers students of three age groups from 3 to 10th grade, all tasks are closed answer type items. All schools are provided by detailed comparative reports. A construct of higher order thinking skills was developed, which consists of two main domains: critical thinking and problem solving.
Main findings. Differences between boys and girls results were observed: girls outperformed boys in critical thinking and boys outperformed girls in problem solving domain; grade has an impact on the results in all age groups; observed quite big differences by regions (urbanization level); there are no differences between e-test and paper version.
Keywords: higher order thinking skills, problem solving, critical thinking.