Assessing experimental science in 11 – 18 education

New research directions

Conference report

Wednesday 12 October

THE ROYAL SOCIETY

The Royal Society was able to hold this conference with the kind support of the Gatsby Charitable Foundation.

Executive summary

Experimental science has always been at the heart of teaching and learning in science classrooms. There have been some concerns raised in recent years about the ways in which experimental science has been assessed in England, both in terms of the methods used and how valid these are in terms of science learning and the backwash effect these assessment methods have on the way teachers and students approach practical work in science.

This conference was set up specifically to try and refocus the science and assessment communities on the possibilities there might be in the future to assess experimental science in a way that more closely matches the opportunities that science learning could offer in the modern classroom.

Accordingly, the conference looked to other subject domains, research and experiences in other countries and the potential of technology for messages, practices and potential that might inform science learning and assessment through experimental science. Current context and challenges were illustrated through examples drawn from research into other subject areas, international projects and technological solutions under development.

Evidently approach and resource play heavily in the success of practical assessment, with the impact of teachers being key. Challenges faced include the need to define experimental science and agree a common language, a framework and a roadmap to improve assessment. Application of measurement systems relies on a firm understanding of what is being measured and research is required to define appropriate measurement and comparison techniques, how to validate results of the assessment process, and how to support teachers throughout.

Digital technologies can be used to record contextual assessments and automatically rank students' results for comparative analysis. Benefits of computer-based testing include access to complex process skills, the availability of performance and process data and the provision of straightforward tools to mimic real-life scenarios and experiments. The application of technology drives changes in research, enabling automation of more complex tasks. While this offers many benefits, there may remain resource issues for science assessment.

The comparative judgement approach to assessment relies on variation in responses and recruitment of expert judges, but has been shown to be consistently valid and reliable. In addition, combining summative and formative assessment helps minimise distortion created by the process itself. It was agreed that learning and measuring should go hand-in-hand, and that data use and balance between approaches require particular attention.

Future research might look at how students should learn science and the skills this entails; the validity of teacher assessment (including the need to increase confidence in this by mapping and developing teacher assessment competences and the use of comparative judgement); and the integration of summative and formative assessment. When defining essential skills, experimental science should be looked at holistically and within different contexts without losing sight of the need to inspire the students. It is essential to 'look from the classroom out'.

Introduction

Professor Tom McLeish FRS
Chair of the Royal Society's Education Committee

Science calls on creativity and imagination as deeply as does music or art, so why should all students not practise experimental 'hands on' science? However, overtight scrutiny and formulaic assessment can suck this sort of life out of education. To help overcome these pressures, the Royal Society wants to generate innovative thinking and initiate new avenues of research into the assessment of experimental science in schools.

During my research career, I have found that the most imaginative ideas come from sparks ignited across disciplinary and sub-disciplinary gaps. Using this approach, the Royal Society brought together researchers from a number of fields to discuss research into the assessment of experimental science.

This conference was designed to encourage active debate among participants. We commissioned a number of pre-conference papers to spark new ideas before the day itself. The conference then featured presentations by academics from a variety of disciplines and countries covering:

- Summative assessment in non-science disciplines;
- Tomorrow's world: exploring innovative approaches to assessment;
- Towards new research directions: group discussions.

"During my research career, I have found that the most imaginative ideas come from sparks ignited across disciplinary and sub-disciplinary gaps."

Professor Tom McLeish FRS

Teachers are critical in the formation of the extraordinary community of people that 'make science happen'. Small-group and plenary discussions enabled us to consider the opportunities and dilemmas new approaches may raise for teachers and schools. Increasing the capacity for research collaborations between schools, universities and industry is one way of supporting teachers to invigorate science education.

Keynote address

Professor Jens Dolin
University of Copenhagen

In his keynote address, Professor Jens Dolin compared formative use of assessment with an emphasis on validity to summative use of assessment with an emphasis on reliability. He noted that an overall solution must align the two such that student competences might be summatively assessed in a valid and reliable way, without distorting the everyday, formatively oriented teaching and learning.

When tackling experimental science assessment, Dolin recommended application of the Nordic didactic approach. Collaborative partners must:

1. Clarify the role of experimental work based on an understanding of what science is within the conceptualisation of the nature of science (McComas *et al*. 1998);

2. Establish a valid framework for experimental science, ie a model of experimental competence, including a theoretically and empirically grounded learning progression;

3. Develop an assessment design able to monitor and judge the whole framework and deliver evidence of student learning and levels of attainment in a reliable way. This means:

    a. Tackling 'The situation problem': assessment of competence needs a rich test environment, a social and cultural context, which offers possibilities of performing processes such as inquiries and investigations;

    b. Research assessment time-spans: how long do students need to practise tasks reflecting the intended learning outcome in a valid way?

> "Experimental science should not be seen as an independent activity but as part of an integrated endeavour to develop students' (scientific) understanding of the real world."
>
> **Professor Jens Dolin**

> "Does Jens's solution work only in a context where trust in teachers' assessment by public and policy-makers is high?"
>
> **Plenary discussion question**

The Validation of PISA (Programme for International Student Assessment) research project compared students' PISA scores on specific PISA items with their scores in a socio-cultural context. The results demonstrated that valid assessment of experimental competences requires the assessment to be performed in an authentic context, or else '... test results will overstate the students' actual learning' (Looney 2011).

An ambitious and viable solution will combine the formative and summative use of assessment in order to align accountability and the learning purposes of assessment. Dolin suggested that digital technologies can be integrated to formatively improve learning and also track performance for summative purposes. Teachers can carry out assessment using evidence from ordinary activities, supplemented by evidence from specially devised tasks, typically collected via portfolios. But, he said, any systematic change to teaching and assessment must be based on cooperation between teachers, researchers and policy-makers.

SESSION 1

Summative assessment in non-science disciplines

Chair Dr Christine Harrison, King's College London

The current assessment system looks at experimental science through a small number of written examination questions, but this may not be the right approach. This session looked outside science for insights into how we might assess experimental science. Presentations were followed by small group discussion and a plenary discussion to allow new ideas to be shared.

Digital approaches to authentic performance assessment

Professor Kay Stables, Professor of Design Education, Goldsmiths, University of London, outlined how digital technology is being used to capture and assess performance capabilities in design and technology.

Contextually based tasks are recorded throughout the design process to produce an online portfolio of work as a 'journey' in real-time. Handheld technologies are used to photograph work carried out under exam conditions across 6 morning hours over 2 days. The Adaptive Comparative Judgement (ACJ) engine algorithm applies quality assessment based on Thurstone's Law of comparative judgement (Thurstone 1927) and developed between 2004 and 2009 by the e-scape project team (Technology Education Research Unit at Goldsmiths University of London, and Digital Assess), which dynamically generates a rank of entries as judges, who could be based around the world, make detailed comparative judgements of multiple pairs of portfolios. The system produces high levels of reliability and validity.

This session outlined a pragmatic and progressive way of collecting evidence of learning throughout a course, producing an electronic portfolio of work that illustrated the quality of work produced and also reflections on the process and progress of students.

Potential issues created by unilinear tracking

Professor Martin Fautley from Birmingham City University shared insight from measurement and judgement in music assessment.

Delivered as a highly isolated subject across most schools, with an open curriculum, music assessment relies on criterion-referencing and differentiated grading assessed by teachers, and often lacks moderation. This situation can lead to pseudo-quantitative measures to compare results between schools. For example, comparative assessment grids using a three-point scale ('not yet able to', 'able to', 'exceeds'), specific to each project, yet applying only to that project (Fautley & Daubney 2015). Senior leadership often requires teachers to implement tracking systems to 'prove' where their students sit on a line. As a result, in some schools only unilinear progression is permitted with minimum scores having implications for curricula and schemes of work, and in many cases teachers come to believe that attainment is no longer important, only progression.

This session highlighted some of the problems that can besiege a subject when the demands of accountability clash with the desired assessment approach within the subject.

> "Assessment needs to be the servant of learning."
>
> **Dr Christine Harrison**

Practical assessment and audio-visual exam submissions

Jennifer Stafford-Brown, Stafford-Brown Consultancy Ltd, provided an overview of assessment for GCSE Physical Education (PE), which is practically assessed in schools (60% practical assessment, 10% course work, 30% written exam) with teachers choosing when to assess students based on when the student is performing at his or her best. While inclusive and accessible, narrow grade boundaries and top-end bunching suggest a bias in teacher assessment, favouring students, and call into question the reliability of this approach. The external quality assurance process did not look at the entire population and inconsistencies between sports presented difficulties in assessment.

A reformed approach has been introduced as of September 2016 for GCSE PE (60% written, 10% course work, 30% practical). In addition, a new vocational qualification is due to launch in 2017 which for the first time will include an externally set, externally marked, 10 minute practical assessment where the student is recorded delivering a sports leadership presentation. Existing vocational practical assessment already uses this audio-visual evidence approach and issues raised included quality of footage, technology costs and staffing concerns. However, after 5 years the situation has improved dramatically.

This session looked at the compromise that PE has made to balance the demands of reliability and validity and how the use of video as a means of capturing the practical aspect of the subject has been successfully achieved.

> "Is it pressure from head teachers that drives grade inflation in teacher assessments?"
>
> **Plenary discussion question**

A lesson in defining the aims of education and assessment

Dr Shaun Helman, from the Transport Research Laboratory, presented work on the safety of young and novice drivers. Results from decades of research show that both age and inexperience play a part in the inflated crash risk of newly licensed drivers. The older someone is when they begin solo driving, the safer they are. In addition, newly qualified drivers of all ages get safer as they accumulate on-road experience.

Helman noted what might be considered a reflection on the validity of the main instrument we use in the UK to establish someone's suitability for solo driving (the practical driving test): those who most readily pass it (young males) are also the group that most readily find themselves in serious injury accidents. However, some findings also demonstrate that the practical test (and theory test) can have benefits for safety. For example, the sub-group of people who pass first time, when age, experience and gender are controlled for, have fewer crashes per mile driven (Sexton & Grayson 2009).

The introduction of hazard perception testing into the theory test in 2002 has also led to safety improvements (Wells *et al*. 2008).

Other evidence shows that pre-test practise on road can be protective of later crash risk, and there is a new test under development (and evaluation). The research trial, to be completed in 2017, will assess if the test changes the way people learn and if it impacts collision rates.

This session was important because it indicated that we need to think about assessment from a consequential viewpoint. Are the assessments selecting people that are likely to be successful in those activities post-assessment?

> "We need to find out what they are learning (post-test) and learn how to teach it to them (pre-test)."
>
> **Dr Shaun Helman**

Group discussion

What was the most striking thing you heard?

Discussion among several groups centred around the use of data, for both formative and summative assessment, application of data to ranking and progression, and the need to differentiate grades while remaining fair and accurate in our measurement approach. Current PE assessment highlights the issue of grade bunching and 'flight path syndrome' (an over-emphasis on measurement rather than guidance for progress) was expressed as a concern.

The driving test presentation was cited as a good example of authentic research and an assessment approach that steers learning correctly.

Another focus of discussion was around the role of teachers and the need to ensure they are not over-pressured by external forces. While teacher assessment can be unconsciously biased, teachers can be trusted and support is needed to help develop their ability to manage bias while remaining integral to the assessment process. Comparative testing was cited as having the potential to employ the strength of teachers working collectively.

Group discussion

What parallels did you see with your own research interests?

It was almost universally agreed that the main challenge is to determine which STEM skills and competences are required by society and what progression in practical science skills looks like. We must redefine what we want to measure and then how to measure it. Clarity is needed around desired educational outcomes, as well as further research around how to match those desires with what is possible.

Issues remain around how to present practical science to maximise a student's knowledge retention, as well as how to assess ability to reproduce scientific process. Science is rarely presented in written forms in the real world and assessment should measure a student's understanding of evidence. It was noted that Ofqual has carried out research on how well skills are retained by university entrants from A-level, and specific challenges discussed around the implementation of assessment included the cost of direct assessment and how summative assessment for a group would have an emphasis on group work. Potential approaches discussed included student curation to reduce assessable content and the addition of merits to extend grading beyond simply a pass or fail.

SESSION 2

Tomorrow's world: exploring innovative approaches to assessment

Chair Dr Anna Walshe, National Council for Curriculum and Assessment, Dublin

This session looked at examples of assessments in STEM and how they might be adapted. Two pre-conference papers acknowledged the need for assessment to evolve through collaborative discussions among researchers, policy-makers and teachers, and a further paper provided thoughts on the use of digital technologies for improving experimental science assessment.

Online practical component for 'Validation of Assessment for Learning & Individual Development': a New South Wales (NSW) Department of Education innovation

Joanne Sim and Annalies van Westenbrugge from the NSW Department of Education, School Performance and Improvement team introduced via a video link an innovative online practical component developed as part of the state wide interactive multimedia diagnostic science tests that are delivered annually to students across NSW.

The 'Validation of Assessment for Learning and Individual Development' (VALID) programme builds upon the Essential Secondary Science Assessment begun in 2005. It now provides online end-of-stage assessments for the science curriculum across years 6, 8 and 10.[1] The programme is underpinned by three major components: an assessment framework, NSW's science K-10 syllabus, and the educational taxonomy known as Structure of Observed Learning Outcomes. The tests incorporate various multimedia assets to provide items sets, each of which has stimulus material and contextually linked test items. The composition of the test items is drawn 50% from the skills domain of the syllabus and 50% from the knowledge and understanding domain of the syllabus.

Online delivery allows for the inclusion of a range of multiple choice and short response item types. The test platform automatically captures and records all responses for assessment. The short responses are also automatically scored. The tests also include open-ended, extended response tasks that enable students to demonstrate higher-level thinking and use of metalanguage, as well as highlight their misconceptions and misunderstandings. The affective domain is assessed by including survey questions drawn from the values and attitudes outcomes included in the science syllabus. Teachers are provided with 5 hours of registered training to reinforce staff capability in making consistent judgements against syllabus standards. Schools, parents and students receive the full analysis of students' achievement by accessing their data through the Department's data analysis platform.

Those wishing to experience the VALID system can do so at **bit.ly/validlondon**

This session presented a means of evidence collection in which a broad range of evidence can be collected.

1. Year 6 corresponds to students turning 12 that year (the final year of primary/elementary school in NSW). Year's 8 and 10 correspond, respectively, to students turning 14 and 16 in these years.

Assessment in a knowledge economy: new approaches and learning analytics

Professor Patrick Griffin, University of Melbourne, introduced the ATC21S Test Menu, which investigates the assessment of collaborative problem solving.

The Laughing Clowns task is an example of an online task that has two remote players attempting to place 12 balls into the clown's mouth and is used to teach collaboration in graduate schools. The players must work out a code system to enable cooperation and communicate that code across the network in order to complete the task.

Online delivery enables every action and chat event to be recorded. Algorithms can then be used to find and interpret these data to build a skills progression (see figure 1). Students and teachers can log on and see where they lie on the progression. Results are used to assess scientific knowledge, ability to collaborate, communication and problem analysis and solving skills, as well as social and cognitive skills. Tested in six countries, these assessments are highly engaging.

Those wishing to experience the ATC21S Test Menu system can do so at **education.unimelb.edu.au/about_us/educational-software-suite**

This session provided insights into assessing some of the transferable skills that are easily developed in science.

FIGURE 1

Skills level progression built by algorithms from data gathered by the ATC21S Test Menu.

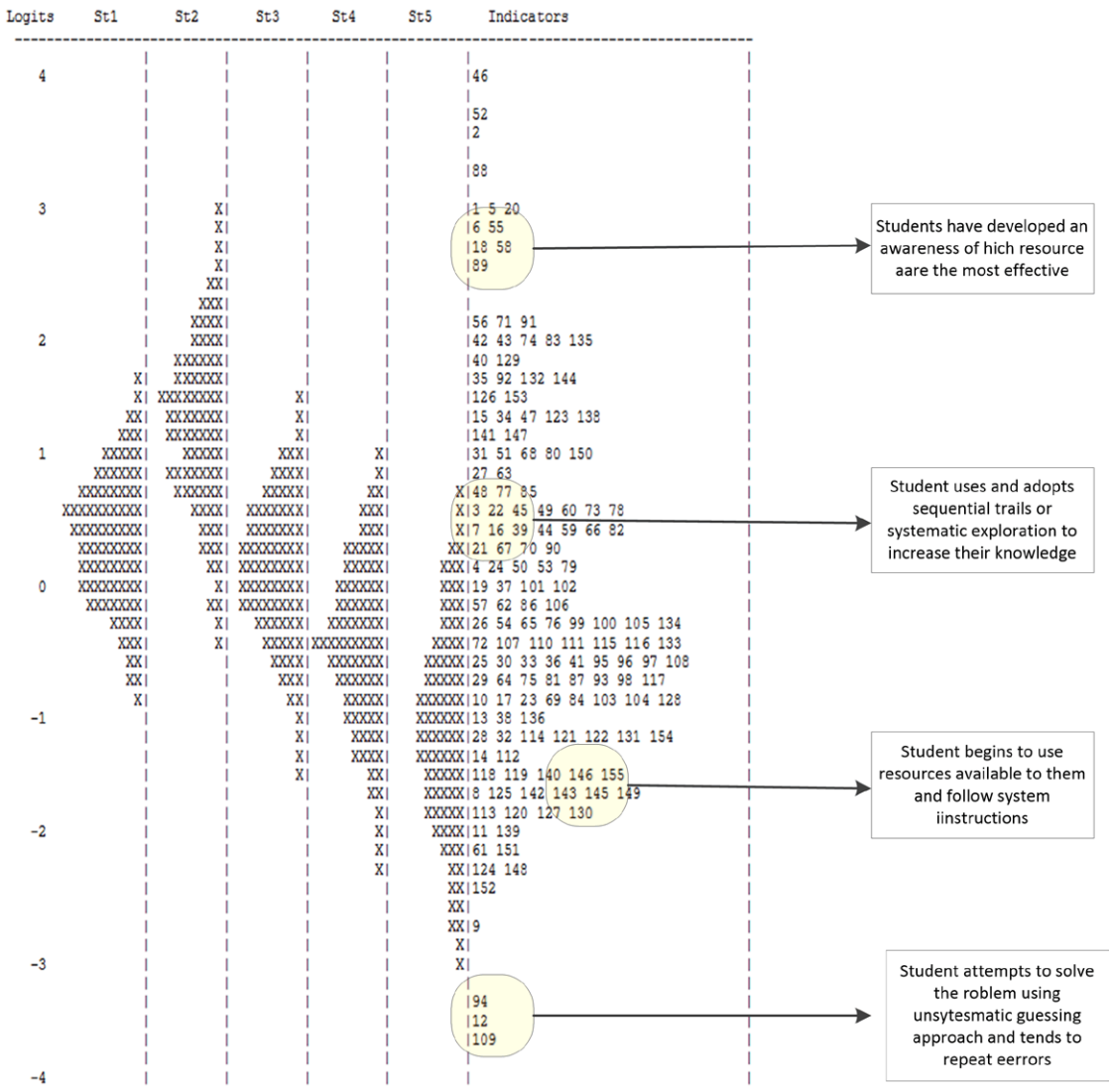


Image reproduced with permission from Professor Patrick Griffin.

Computer-based testing of complex problem solving

Dr Ronny Scherer, University of Oslo, outlined current innovations and challenges in computer-based testing (CBT) by looking at complex problem solving activities and how students shift from one strategy to another as they move through the problem-solving process (figure 2). Minimal Complex Systems (MCS) are computer-based

FIGURE 2

The problem-solving process.

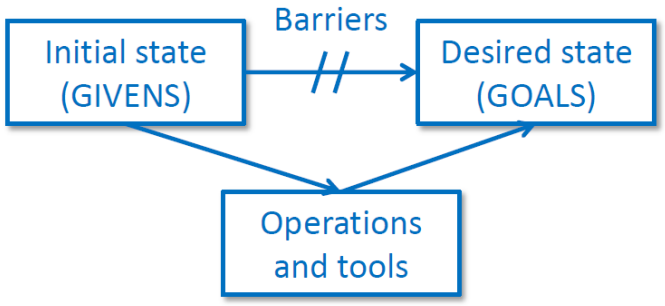


Image reproduced with permission from Dr Ronny Scherer.

assessment tools that set tasks within limited systems with limited variables, eg a climate controller that can influence temperature and humidity. Students must interact with these systems by systematically varying variables to see the effects and generate knowledge about how variables are interrelated. MCS are used to track a student's ability to build a mental model required to solve a given problem. Moreover, with the help of log file data, changes in strategic behaviour or response times can be examined.

Scherer believes the challenges of CBT include the complex measurement and structural models needed to describe the development of skills, selected aspects of accessible constructs and the often unclear meaning of constructs, such as response times, that require validation. Benefits of CBT include access to complex process skills, the availability of performance and process data and the provision of straightforward tools to mimic real-life scenarios and experiments.

Those wishing to experience the MCS can do so at the OECD's PISA 2012 webpage
oecd.org/pisa/test/testquestions/question3

This session provided a view of how technology might be developed to assess complex skills in the future.

Roadmaps to help develop assessments for learning progressions in science

Professor Mark Wilson, University of California, Berkeley, introduced two examples of the Berkeley Evaluation and Assessment Research Center (BEAR) Assessment System (Wilson & Sloane 2000) and its application to Learning Progressions to demonstrate the need for careful application of any measurement system.

When analysing a 'Structure of matter' learning progression, Wilson found that instead of a set of steps going up from bottom-left to top-right, the results were highly irregular (figure 3). However, altering the construct by (a) splitting it into three sub-constructs (strands A – C, figure 3), and (b) therefore re-ordering some items resulted in the sort of step-wise progression originally expected.

Wilson has also researched how students apply the Toulmin Argumentation Model, connecting claim, warrant and evidence through argument. Results demonstrate, somewhat paradoxically, that for students it is relatively harder to critique somebody else's argument than to construct an argument oneself.

According to Wilson, specification of learning progressions demands a greater focus on 'what' is developing for curriculum, assessment and instruction, but the rewards are rich for assessment, professional development, and hence for students. Psychometricians must apply uni- and multidimensional models to interrogate the data appropriately to represent and model 'links' across dimensions.

This session outlined the difficulty of mapping and assessing progression.

> "We did not know what we were measuring until we measured it. Measurement and testing must happen simultaneously."
>
> **Professor Mark Wilson**

FIGURE 3

Initial progression results (top) compared with results split into three enquiries by adding a new dimension to the construct (bottom).

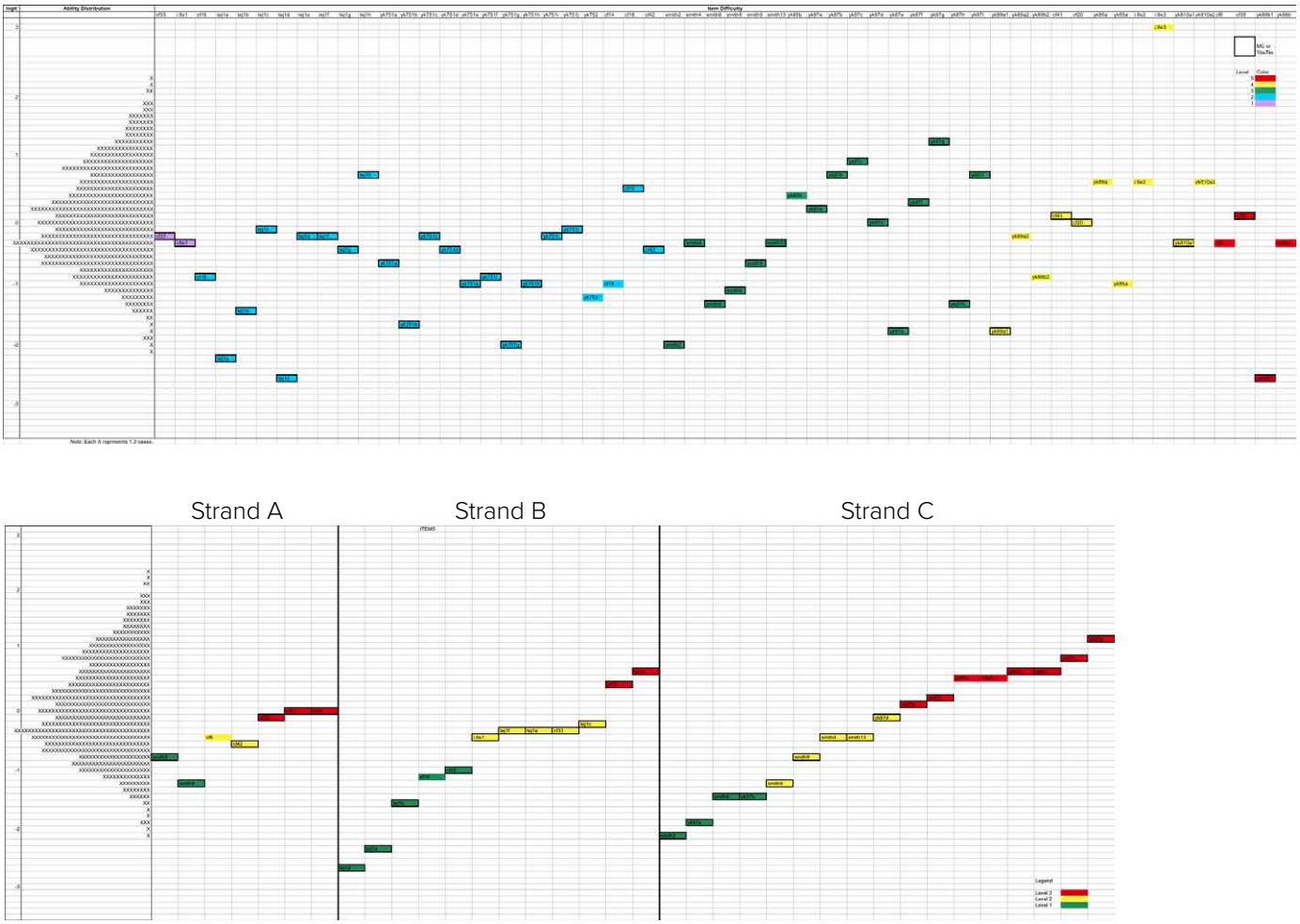


Image reproduced with permission from Professor Mark Wilson.

Comparative judgement for robust assessment

Dr Ian Jones, University of Loughborough, described how comparative judgement is used to assess student responses to open questions aimed at testing conceptual knowledge in mathematics, such as: 'What is an equation? Give examples of how equations can be useful' (see figure 4).

The comparative judgement approach places two examples side by side and asks a judge to compare the responses and select the student with the better understanding of the subject. The method delivers binary decision data, which can then be modelled and scored for each student.

Comparative judgement requires a widespread variety of answers. Evidence for this method is founded on the Law of comparative judgement (Thurstone 1927). It can be applied to a range of topics (for example, fractions, calculus, statistics, geometry) and contexts and is used to assess conceptual understanding and problem solving.

Practically, the approach requires recruitment of expert judges such as school teachers or PhD students, to gather between 5 and 12 judgements per script. Giving each script the same number of judgements feeds into validity and reliability. It has been shown to be consistently valid by standard quantitative and qualitative evaluation methods, as well as internally and externally reliable (Jones & Alcock 2014; Jones & Inglis 2015; Jones & Wheadon 2015; Bisson *et al*. 2016; Jones & Karadeniz 2016).

This session looked at an alternative way of assessing a cohort and its value in developing better understanding of what quality work looked like.

FIGURE 4

Sample student responses presented for comparative judgement.

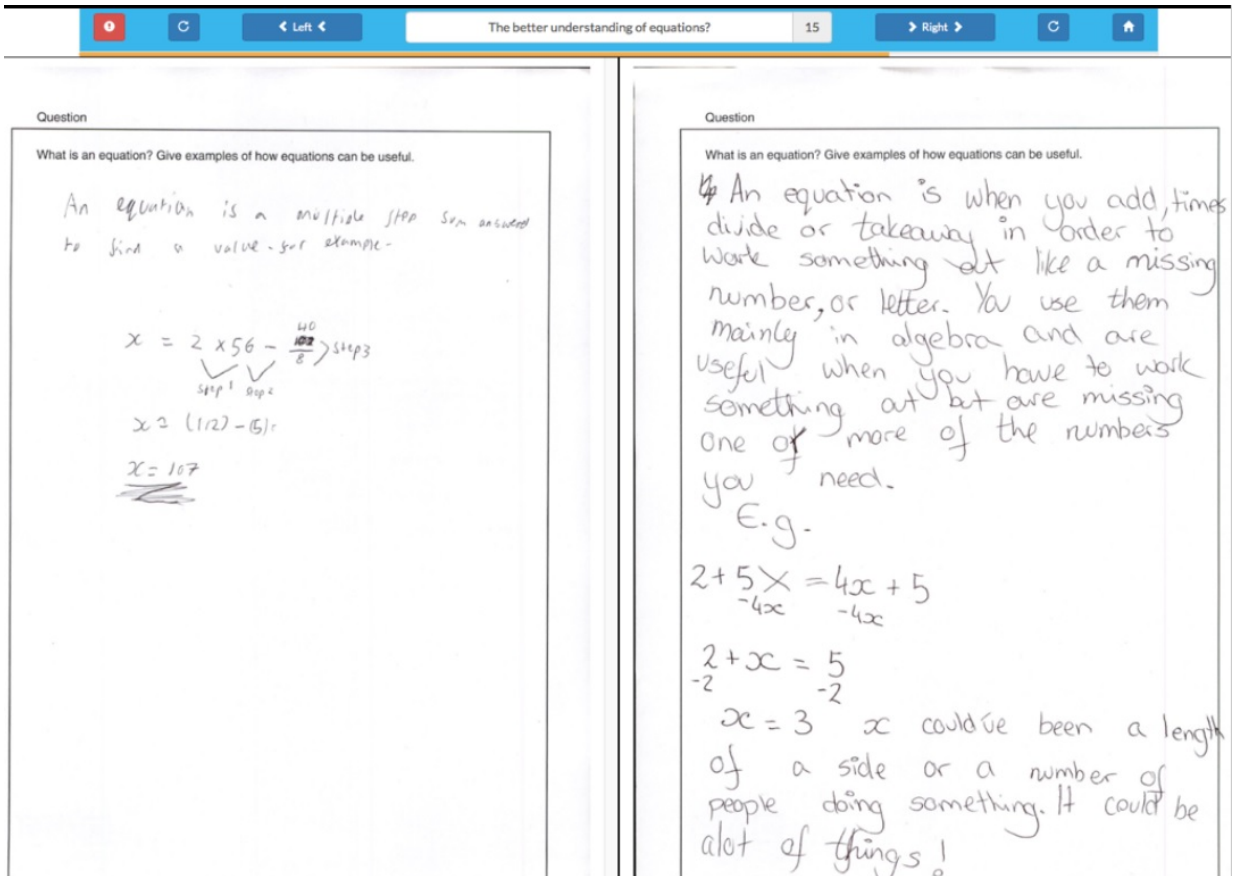


Image reproduced with permission from Dr Ian Jones.

Sharing standards: how comparative judgement can help

Dr Chris Wheadon, founder of nomoremarking.com, illustrated the reliability of comparative judgement by asking the audience to judge two samples of Year 6 writing. As is typically the case, audience opinion was split 80:20, with the 80% choice considered the correct one.

Teacher-based assessment of Key Stage 2 (KS2) writing moderation was introduced in 2010 and, while reformed for 2015 – 16 as part of wider curriculum change, the system remains based on a list of criteria which research demonstrates to be consistent within schools, but not between schools.

A computer-assisted comparative judgement approach reflects traditional moderation principles and is much quicker and less taxing for assessors. Students' responses are scanned and distributed across multiple schools, allowing all teachers to judge all students.

Wheadon's KS2 trials of this approach resulted in reliability scores of over 0.85, indicating a high degree of agreement amongst judges. Ten teachers could judge a year group of 60 portfolios in 30 minutes, with portfolios taking longer to assess than individual pieces.

The current nomoremarking.com system will be improved in future using bar-coded sheets for student responses and automated script sharing between a target of 250 schools.

This session, like the previous one, built on the ideas of comparative judgement and its use in developing understanding of quality within a subject context.

> "A valid process, comparative judgment allows you to reward real quality."
>
> **Dr Chris Wheadon**

FIGURE 5

Sample student responses presented for comparative judgement.

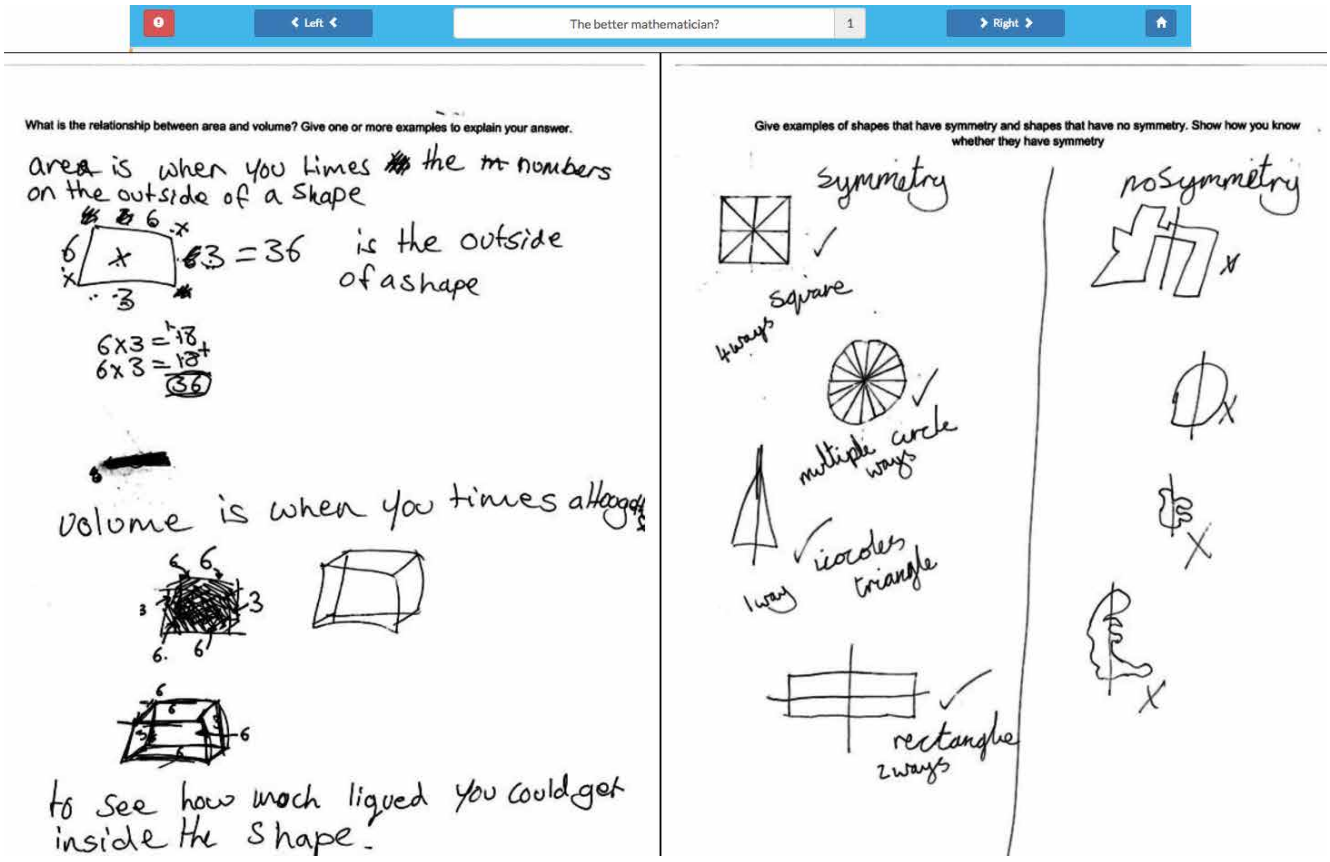


Image reproduced with permission from Dr Chris Wheadon.

Collecting evidence of inquiry learning in the science classroom

Dr Eilish McLoughlin, Dublin City University, looks at how Irish teachers can be supported through changes to classroom assessment practices that recognise and value skills and competences that may be developed in secondary school.

Strategies employed to assess science learning in the classroom involve using inquiry as an active learning pedagogy. Teachers are given an opportunity to gain experience of learning new content knowledge through an inquiry approach requiring them to draw on prior experience and engage in peer discussions. Inquiry approaches can range from guided (learners completing a structured worksheet) to learners completing an open inquiry investigation. Teachers are then asked to reflect on what learning has occurred. In this way, we recognise that teachers' assessment practices are influenced by their beliefs about student learning and their own assessment literacy (Guskey 2002).

Assessment of science learning, ie subject knowledge, skills, competences and attitudes, can be carried out by 'on the fly', structured classroom dialogue and embedded assessments, and teachers are facilitated to realise what and when assessment opportunities are possible (figure 6).

> "If we want a more teacher-led formative approach to assessment, then maybe we need to research ways of improving training in these areas for science teachers."
>
> **Plenary discussion comment**

Collaboration between teachers, researchers and policy-makers is essential in order to ensure that the role of assessment is considered '...looking from the classroom out' and to determine the key objectives of science education:

1. What skills and competences are needed by school leavers for STEM careers and society? Are these identified or developed in school science curricula?
2. What skills and values are recognised by science teachers? Are these skills and values developed or assessed by science teachers?
3. What influences science teachers' assessment practices? What are science teachers' attitudes and beliefs about assessment and student learning?

This session outlined some of the new approaches to assessing experimental science in European classrooms and the support and training teachers need in order to conduct them.

> "We have seen through European studies that if we can provide teachers with appropriate support, they can adopt new classroom practices. Time is always an issue and sustained collaboration is central to affecting this change."
>
> **Dr Eilish McLoughlin**

FIGURE 6

Teachers' role in the assessment of learning.

Image reproduced with permission from Dr Eilish McLoughlin.

SESSION 3

Towards new research directions: group discussions and comments

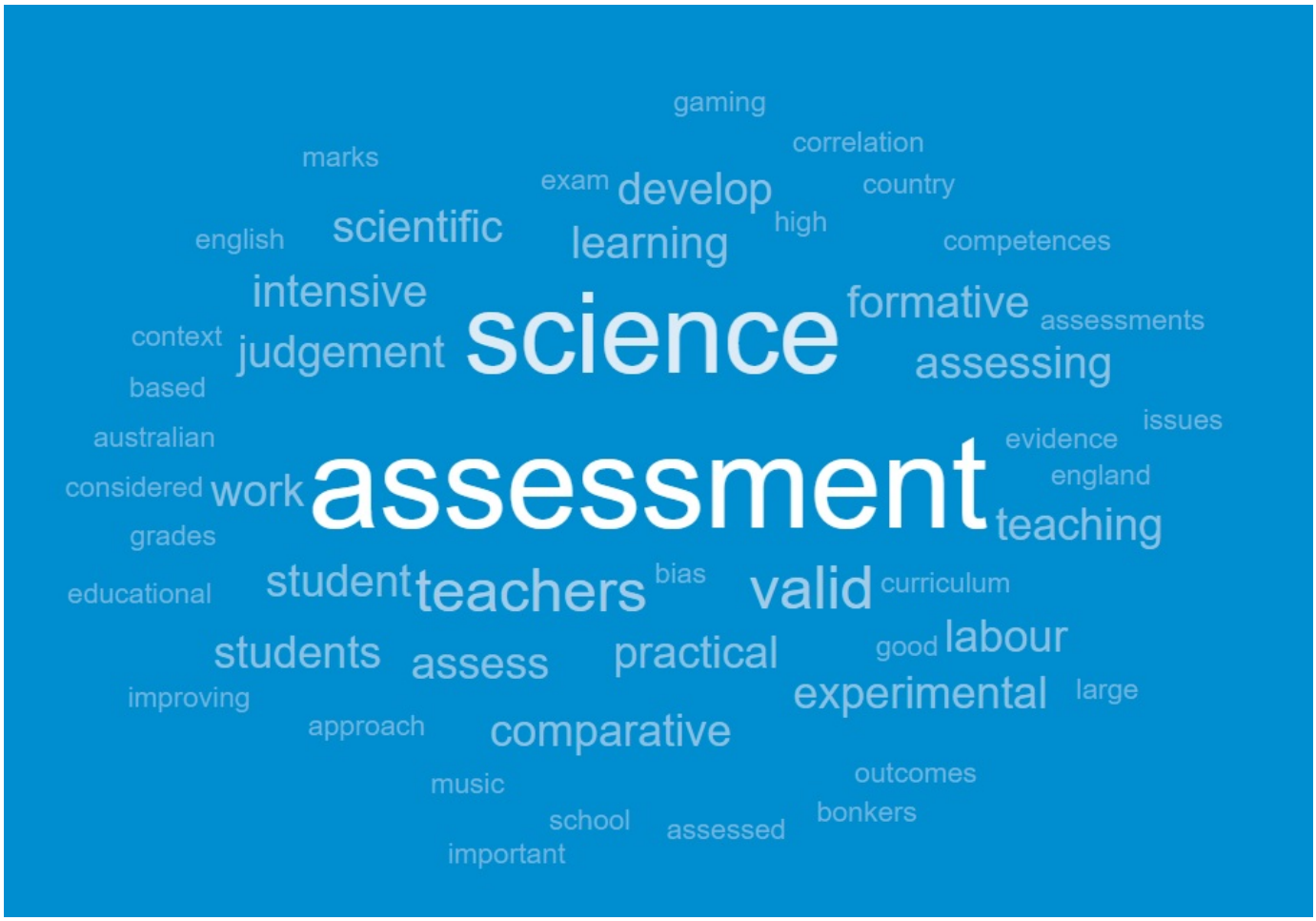
Chair Professor Robert Coe, University of Durham

Each set of talks was followed by a plenary session where audience members were asked to split into small groups and discuss a list of questions, recording their thoughts on flipcharts. Participants' thoughts and questions were also gathered using Slido (see figure 7). Conference attendees were asked to submit ideas on potential research questions in both online and offline discussions.

The plenary session then brought all components of the conference together, and asked participants to consider avenues that could be fruitful for future research on assessing experimental science in schools.

FIGURE 7

Slido-generated word cloud visually representing word frequency throughout the conference.



What is 'experimental science' and what constitutes measurement of practical competence?

Questions and discussion implied a consensus for the science community to follow Professor Dolin's recommendation to develop a framework of competence in experimental science and a roadmap from which to tackle these issues. It was postulated that the current lack of understanding around what should, or could, be measured suggested a need to stop tweaking the current system to allow time for research that might answer such fundamental questions. However, it was also noted that the act of measuring is the act of learning, so assessment and learning should go hand in hand.

One plenary group believed that one of the most important foci of funded future research should be to 'define the domains of what we are actually measuring'.

What should we be assessing and when should we not assess at all?

It was suggested that assessments only measure part of what the student is learning at school, and that non-assessable goals can get lost, so perhaps 20% of learning time should be assigned to non-assessed aspects of curricula. Further debate is required to decide when assessment should focus on the way students learn science, the process they use to learn it, the skills and outcomes they learn, or the progression of learning over a certain period. It was, however, noted that it was necessary to avoid 'flight path syndrome' (an over-emphasis on measurement rather than guidance for progress), that measurement is useful only until the costs of distortion outweigh the benefits of assessment, and that the integration of summative and formative assessment helps minimise distortion.

> "When you focus on something you automatically distort it – the Heisenberg Principle for assessment."
>
> **Plenary discussion comment**

Integrating formative and summative assessment

Much of the plenary discussion focused on summative and formative assessment and the need to combine and integrate the two. Experts in the room believed that how data are used is critical to this process, but further understanding of how to balance the two types of assessment is also needed.

It was noted that, 'Unless you create summative assessment in concert with good diagnostic assessments, the summative assessments are almost always biased'. And that: 'We should reclaim summative assessment, in a positive way. Students like summative assessment. It is fit for purpose. We should also reclaim teacher assessment. If so, we can assess a more diverse curriculum in a reliable fashion and more efficiently than the current external process we use today.'

> "Progress is a ladder to climb; this is the view of teachers."
>
> **Slido comment**

What do we want students to experience and achieve when carrying out experimental science?

Discussions called for agreement around which skills experimental science should aim to deliver and the need for a common approach to practical science. Do we want to see the same experimental science outcomes in biology, chemistry and physics? Employability requires a wide set of skills and one reason we value experimental science is to develop curiosity. To tackle this, it was recommended that experimental science should be looked at holistically and within different contexts without losing sight of the need to inspire students.

> "Can assessment of experimental science be valid without being cross curricular – must require competences that transcend the scientific corner of the curriculum."
>
> **Slido comment**

"All assessment both measures and incentivises, so must be designed accordingly."

Slido comment

Possible approaches included looking at what practical steps professional scientists take to solve problems and gather evidence, rather than looking at the evidence they find, encouraging autonomous learning: 'learning to learn'. The Singapore Quality Award (SQA) model (Ng 2003) was cited as an example of how wider skills can be encouraged within schools and it was noted that valued skills must first be fostered in teachers.

One tested approach is the 'exposure to different skills' approach, available for some time in Ireland, which is done well by many teachers (CCEA 2009). However, it can also result in a significant number of schools completing a 'circus of experiments' without having covered any theory. So, if 'exposure' is the goal, it must be linked to the curriculum.

Valuing students' perspectives

Students are objective in their perspectives. For young people science is simply one subject of a multiple set of subjects they study and each student will navigate their own experience. Choices will be made and individuals will approach their learning in different ways.

It was suggested that peer assessment, with peers seeing examples of one or two grade points ahead (Jones & Alcock 2014; Jones & Wheadon 2015), might provide insight into how individuals might improve and that student curation might produce more manageable amount of assessable content.

One key research topic would be to look at issues faced by students who are non-native speakers and how experimental science could be more inclusive.

Teacher support and continued professional development (CPD)

The role of the teacher was a critical issue, discussed on many levels. If we want a more teacher-led approach to assessment, it is important to research ways of supporting in-service teachers in assessment design, improving teacher training and finding new ways to boost teachers' self-efficacy and self-confidence in the application of assessment methods. While investigating the benefits of extending assessment approaches to align with different teaching strategies, research should include the potential reconstruction and development of school-level assessment and how to avoid league table pressure on teachers.

"In England, we use exam results to assess teachers. This compromises reliability and validity."[2]

Slido comment

It was recommended that assessment boards provide professional development to share practices for experimental science, support teacher assessment validity and build confidence. One area of future funded research might investigate if comparative judgement can improve teacher confidence in assessment.

Teachers need support and training in assessment literacy, practices and evaluation. Training teachers, from design to moderation of assessment, and how to support them better was noted by several plenary groups as being important for future research.

"Teachers should not feel over-pressured by external forces; teachers should be trusted."

Plenary discussion comment

2. The UK Government does not use exam results to assess teachers, nor does Ofsted, but individual teachers may do so.

Technology and its application

The application of technology means that research methods have changed when compared with 10 years ago; we can now link conceptual ideas, learning progression and data. Cloud-based solutions enable pragmatic collection and storage of e-portfolios, audio- and video-based evidence of practical competences, as well as automated dissemination and analysis to support computer-based assessment and assessment techniques such as comparative judgement.

"Technology; master or servant?"

Slido question

In order to encourage thinking about the future impact of technology on assessment, during lunch and subsequent breaks, participants were invited to see demonstrations of two digital technologies: Labdog (a Web application developed at the University of Southampton to facilitate teaching and real-time or retrospective assessment in the teaching laboratory – see **edtechandchem.ghost.io/reintroducing-labdog**) and Labster (see **labster.com**; a virtual lab that enables students to carry out experiments and have their performance tracked and graded using gamification methods). These demonstrations added new perspectives to how education technology can bring together student motivation, learning effectiveness and assessment in a process where virtual and physical learning environments will supplement each other.

While it was generally agreed that technology offers many potential benefits, it was also noted that some methods, such as collection of video footage, might have resource issues for the large cohorts that need to be assessed in science. One participant also asked if comparative judgement assessment might be too labour intensive to be a viable means of assessing student lab portfolios, but was informed that those with experience of the method felt it was less labour intensive than other methods.

Several of the key questions highlighted by plenary groups involved further research in the use of technology for assessing experimental science:

- How can we harness the interface between human skills and systematic machines to achieve the best of both?
- How can digital portfolios be used for science moderation?
- Would e-portfolios be feasible for science assessment? Can this approach handle the different ways of measuring the constructs required?

Who should be solving these issues? Collaboration and policy

Open discussion raised concerns as to whether the responsibility for solving some of these issues lay with government, rather than the assessment community, and if it was worth waiting to find out if the current approach will work or not. Will changes currently being implemented change students' behaviours in schools rather than assess their skills?

"Should an assessment community have to solve what might be considered as a consequential validity issue which has been created by the Government?"

Slido question

These questions provoked strong opinions and it was suggested that, while there is work to be done to strengthen the current assessment community, these matters should be tackled by that community, in collaboration with related communities (industry, teachers, teaching unions and policy-makers) to ensure the right questions are asked and answered, and to help guide improvements in policy. International research collaboration would bring shared benefits from seeing how different countries address challenges.

It was also noted that current issues of consequential validity (using assessment to drive instruction) are as much a fault of the assessment community as government and should not be solved without input from others. Is the consequential validity issue worth considering as a linear programme, that is can we measure something and incentivise behaviour in some way? Would that help to create a framework if it were recognised from the very beginning?

"There is evidence to be collected on the impacts of recent curriculum reform, but how can we ensure the next curriculum reform is evidence-based?"

Slido question

Plenary group discussion raised an exercise in collaborative thinking to bring everything together as being another important focus for future funded research and noted that funding associations might come together to carry out a systematic review of everything already done, produce a gap analysis and fund research projects to fill the gaps.

While some factions were keen to monitor how recent changes in assessment of practical skills affect learning in science, the majority of the participants called for a more radical approach to prepare us for the future.

"We are seeking evolution rather than a revolution at this point. We are not saying that the current system is wrong, we want to focus on moving things forward in the future rather than trying to solve problems now."

Dr Chris Harrison

Funding and other potential research questions

In addition to matters already discussed, other research suggestions included:

- Geography: what does it look like when changes and aspects such as assessment of fieldwork are reintroduced?
- Further study into teacher assessment competency. Do teachers confidence and beliefs change throughout the project?
- Can we crowd source assessment questions, gather feedback from other teachers and create collaboration between schools?
- How does a teacher's subject knowledge affect their use of formative assessment?
- What is happening at initial teacher education (pre-service) to prepare teachers to use formative assessment for experimental science?
- Research on progression paths for working scientifically.
- What is the impact of direct versus indirect assessment of skills and competences?
- A comparative resolution between what science students do/think when carrying out a task compared to 'real' scientists.
- Looking at how teachers develop assessment capabilities – longitudinal study (more detail).

Closing remarks

The Organising Committee felt strongly that discussions should be continued beyond this conference and taken forward to affect what goes on with teaching and learning science in schools. They back the aim to 'use assessment within the curriculum as a "servant to learning", as it should be'.

Professor Sir John Holman, President of the Royal Society of Chemistry, observed that while some disagreed with Ofqual's newly implemented solution, we did not have a better solution based on evidence.

> "So, here we are – seeking to fill that gap. We have time to research and develop alternative systems. We have a shared goal of better assessment of practical science, better learning in science and better engagement in science."
>
> **Professor Sir John Holman**

Summary of ideas generated in session 3

1. What would be your dream collaborative project involving one or more of the speakers today?

2. Which research techniques or approaches to assessment would be harder to translate to your own research interests? Why?

3. Which of the presentations today has the most promise for developing more valid (or reliable) assessments of experimental science? Why?

The above three questions elicited the following comments:

- Fundamental question: what is meant by experimental or practical science?
- Define the learning outcomes of practical work.
- How does a teacher's subject knowledge affect their use of formative assessment?
- What is happening at ITE (pre-service) to prepare teachers to use formative assessment for experimental science?
- Research on progression paths for working scientifically.
- We need to define a domain before moving forward; what is the science curriculum supposed to be delivering?
- Behavioural insight could be brought to bear on assessment – to predict unintended consequences.
-
- Research methods have changed when compared with 10 years ago; we can now link conceptual ideas, learning progression and data.
- We need policy-makers, teachers and teaching unions involved in the research process from the start to ensure we are answering their questions.
- International collaboration on research to see how different countries address challenges.
- Cloud-based, pragmatic collection, storage, use of e-portfolios, tools and techniques.
- Assessment boards should provide CPD to share practices for experimental science, support teacher assessment validity and build confidence.
- Peer assessment with peers seeing examples of one or two grade points ahead so they know how to improve.
- Questions around students who are non-native speakers and how experimental science could be more universal?
- Comparative judgement offers a possible method of making assessments in hard to categorise disciplines.
- Online assessment tools show great promise.
- Comparative judgement captures expertise in the environment.

Acknowledgements

The Organising Committee

Dr Christine Harrison (Chair), King's College London

Dr Ann Childs, University of Oxford

Professor Robert Coe, University of Durham

Dr Ian Jones, Loughborough University

Dr Tim Leunig, Department for Education and London School of Economics

Professor Paul Newton, Ofqual and University of Durham

Dr Anna Walshe, National Council for Curriculum and Assessment, Ireland

Observers

Dr Mat Hickman, The Wellcome Trust

Cheryl Lloyd, Nuffield Foundation

Ginny Page, The Gatsby Charitable Foundation

Opening remarks

Professor Tom McLeish FRS, Chair of the Royal Society's Education Committee

Keynote address

Professor Jens Dolin, University of Copenhagen

Closing remarks

Professor Sir John Holman, President, Royal Society of Chemistry

Speakers

Professor Martin Fautley, Birmingham City University

Professor Patrick Griffin, University of Melbourne

Dr Shaun Helman, Transport Research Authority

Dr Ian Jones, Loughborough University

Dr Hilary Leevers, Head of Education, The Wellcome Trust

Dr Eilish McLoughlin, Dublin City University

Dr Ronny Scherer, University of Oslo

Ms Joanne Sim, Department of Education, New South Wales

Professor Kay Stables, Professor of Design Education, Goldsmiths, University of London

Jennifer Stafford-Brown, Stafford-Brown Consultancy Ltd

Ms Annalies van Westenbrugge, Department of Education, New South Wales

Dr Chris Wheadon, No More Marking

Professor Mark Wilson, University of Berkeley

The Royal Society gratefully acknowledges the generosity of the Gatsby Charitable Foundation in supporting the conference and this report, and the Wellcome Trust for the use of its conferencing facilities.

References and further reading

Bisson, M-J, Gilmore, C, Inglis, M & Jones, I 2016 Measuring conceptual understanding using comparative judgement. *Int. J. Res. Undergrad. Math. Educ.* **2**, 141 – 164. (See link.springer.com/10.1007/s40753-016-0024-331 October 2016)

CCEA 2009 *Northern Ireland Curriculum. Primary (Irish Medium)*. (See ccea.org.uk/sites/default/files/docs/curriculum_im/area_of_learning/NIC_Primary_IrishMedium.pdf)

Fautley, M & Daubney, A 2015 An assessment and progression framework for music – Secondary. London: Incorporated Society of Musicians (ISM). (See ism.org/nationalcurriculumMFAD)

Guskey, T R 2002 Professional development and teacher change. *Teachers and Teaching* **8**, 381 – 391.

Jones, I & Alcock, L 2014 Peer assessment without assessment criteria. *Studies Higher Educ.* **39**, 1774 – 1787. (See tandfonline.com/doi/abs/10.1080/03075079.2013.821974)

Jones, I & Inglis, M 2015 The problem of assessing problem solving: can comparative judgement help? *Educ. Studies Mathematics* **89**, 337 – 355. (See link.springer.com/10.1007/s10649-015-9607-1)

Jones, I & Wheadon, C 2015 Peer assessment using comparative and absolute judgement *Studies Educ. Eval.* **47**, 93 – 101.

Jones, I & Karadeniz, I 2016 An alternative approach to assessing achievement. In *Proc. 2016 40th Conference of the International Group for the Psychology of Mathematics Education*, Szeged, Hungary, 3 – 7 August 2016. (See dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/21090/1/RR_Jones.pdf)

Looney, J W 2011 Integrating formative and summative assessment: progress toward a seamless system? OECD Education Working Paper No. 58. (See oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=edu/wkp(2011)4&doclanguage=en)

McComas, W F, Almazroa, H & Clough, M P 1998 The nature of science in science education: an introduction. *Science and Educ.* **7**, 511 – 532. (See link.springer.com/10.1023/A:1008642510402)

Ng P T 2003 The Singapore School and the School Excellence Model. *Educ. Res. Policy and Practice* **2**, 27 – 39.

Sexton, B & Grayson, G B 2009 The accident history and behaviours of new drivers who pass their first practical driving test. TRL published report (PPR427). Crowthorne: Transport Research Laboratory. (See roadsafetyobservatory.com/Evidence/Details/10167)

Thurstone, L L 1927 A law of comparative judgment. *Psychol. Rev.* **34**, 273 – 286.

Wells, P, Tong, S, Sexton, B, Grayson, G & Jones, E 2008 *Cohort II: a study of learner and new drivers. Volume 1 – Main report*. Road Safety Research Report no. 81 (See webarchive.nationalarchives.gov.uk/20100513151012/http:/www.dft.gov.uk/pgr/roadsafety/research/rsrr/theme2/cohort2/cohrtiimainreport.pdf)

Wilson, M & Sloane, K 2000 From principles to practice: an embedded assessment system. *Applied Measurement Educ.* **13**, 181 – 208.

The Royal Society is a self-governing Fellowship of many of the world's most distinguished scientists drawn from all areas of science, engineering, and medicine. The Society's fundamental purpose, as it has been since its foundation in 1660, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity.

The Society's strategic priorities emphasise its commitment to the highest quality science, to curiosity-driven research, and to the development and use of science for the benefit of society. These priorities are:

- Promoting science and its benefits
- Recognising excellence in science
- Supporting outstanding science
- Providing scientific advice for policy
- Fostering international and global cooperation
- Education and public engagement

For further information
The Royal Society
6 – 9 Carlton House Terrace
London SW1Y 5AG

T +44 20 7451 2500
W royalsociety.org

Registered Charity No 207043
Issued: February 2017 DES4648