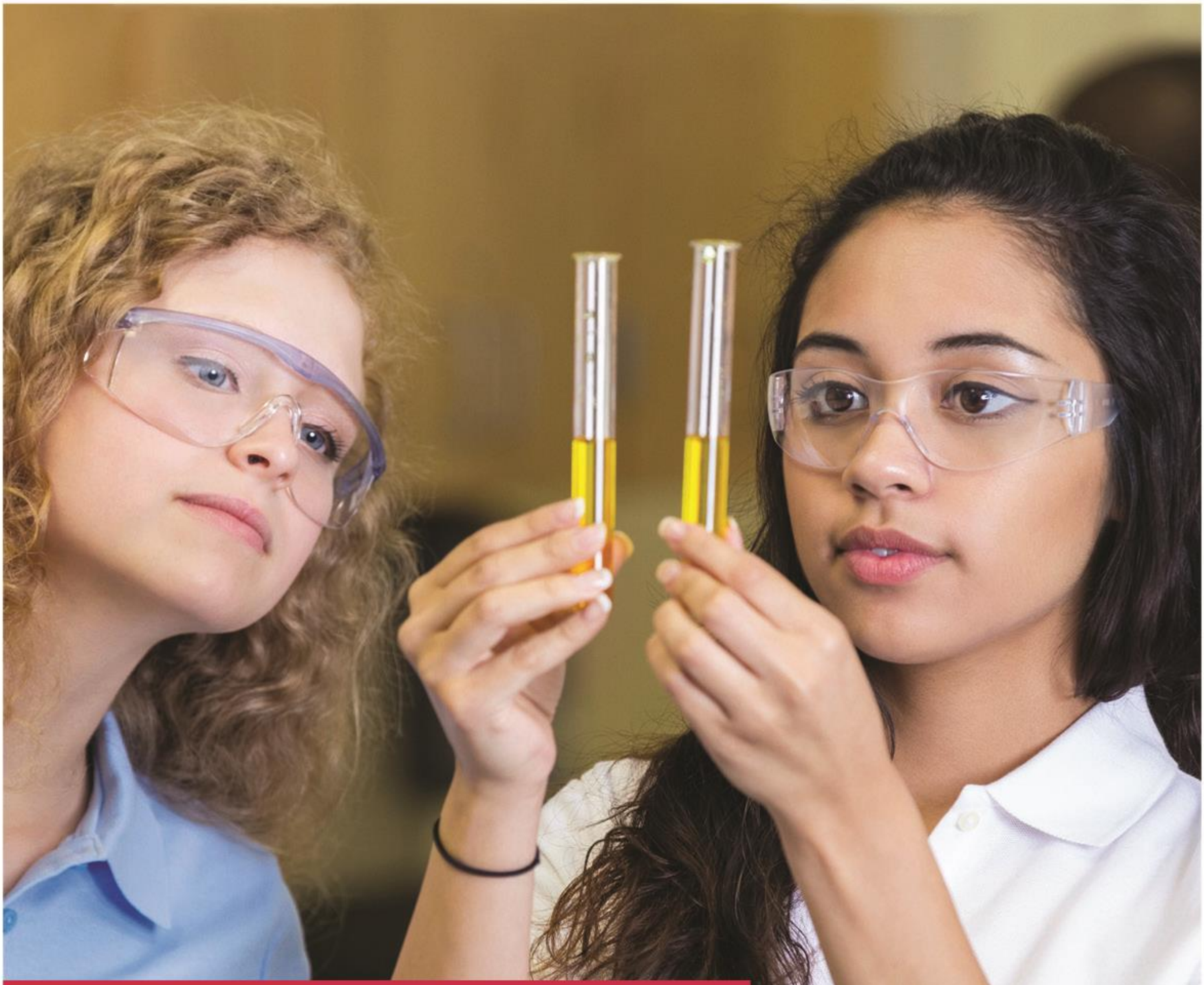

Assessing experimental science in 11 – 18 education: new research directions

Educational research conference

Pre-conference papers

12 October 2016

The Wellcome Trust

THE ROYAL SOCIETY

Cover image

Foreword

On behalf of the Organising Committee, it is my pleasure to welcome you to the Royal Society's conference 'Assessing experimental science in 11–18 education: new research directions'.

We hope that this meeting will provide a rich forum for sharing insights and understanding that inspires innovative approaches to studying assessment of experimental science and, ultimately, improves the way this is assessed.

In order to help make the most of this opportunity, we commissioned the four thought pieces presented here. Together these provide essential background on practical science assessment in the UK and internationally and discuss novel approaches being developed to improve the validity and authenticity of summative assessment in science and language.

Jens Dolin	Improving the assessment of experimental science through evidence-based research: key questions
Jannette Elwood	Reflections on current national examination policy changes at 14–16 and 16–18 across the UK: implications for students' experience of the curriculum and assessment of experimental science
Anthony Green	Directions in language assessment
Michael Seery	Using digital technology to assess experimental science

We hope that the ideas contained in these articles will inform and stimulate your thinking ahead of the meeting, and we look forward to seeing you next Wednesday.

Kind regards

Dr Christine Harrison
Chair, Organising Committee

Contents

Improving the assessment of experimental science through evidence-based research: key questions

Professor Jens Dolin, University of Copenhagen

Summative assessment of experimental science: key problems and the way ahead
Experiments, practical work and lab work have been a core activity in science instruction for centuries and over the last decades hundreds of research papers have been published assessing the effectiveness, student engagement, student learning outcome, etc, of these activities, and yet the present situation still quite problematic. Experimental work is tested in high-stakes tests, via simple practical investigations or using other confined or indirect formats only assessing a limited selection of the important aspects potentially involved in experimental science. This has unfortunate consequences for science education, resulting in lots of resources being wasted on a system that does not improve learning of the many competences potentially learnable through experimental science and needed for a modern society.

The reason for this is a complex mix of political demands, unsatisfying curricula, insufficient use of existing knowledge, unalterable traditions, etc. In this very complex landscape the conference presentation linked to this piece will highlight what we know works, which pitfalls to avoid, what we need to know, and it will give some ideas for future paths of development. The scale of the map must necessarily be large, in accordance with the basic understanding of the need for consistency between the various elements included in summative assessment of experimental science as part of science education. Any consistent assessment system must clarify three closely linked questions:

Why do experimental science? *What are the main goals experimental science shall meet?*
What to do when doing experimental science? *How can we describe experimental science as an integral part of science education?*
How to assess experimental science? *How can we assess student competences in a valid and reliable way – without distorting the everyday teaching and learning?*

Any assessment system answering these questions must have a strict alignment between the three areas! I will cover them one by one with emphasis on the third.

Why do experimental work in science?
The arguments for experimental science have changed over time. From being an independent area of knowledge, involving the ability to act in a lab, it developed into being a support for students' conceptual understanding. During the last decades an ability to perform inquiry processes in more authentic contexts has been vital. Behind these changed explicit goals it has often been a more or less implicit premise that experimental science should be a reproduction of science itself, science as it is practised outside school. The changing arguments for experimental science reflect a continuously evolving (and conflicting!) understanding of the nature of science – of the epistemology of science – which also means that the knowledge *about* science has become part of the learning goals, also for experimental science.

Major research-based curriculum projects, like the US *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*' (NRC 2012), see practical work as a means to help students make links between the real world with its phenomena and events and science's world of models and explanations. The point here is that experimental science is not seen as an independent activity, but as part of an ongoing endeavour to develop students' (scientific) understanding of the real world. Students engaged in argumentation is often accentuated as a key factor in linking the two sides (Osbourne 2015). Any argument for experimental science as part of school science must be based on an understanding of what science is and place experimental science within this conceptualisation.

Clarification of the role of experimental work is the first and most fundamental problem to address.

What to do when doing experimental science?

Shifting the purpose of experimental science, the so-called intended learning outcome, towards inquiry processes in general, and processes in complex, authentic contexts, also outside school, will include generic aspects, like arguing, innovation, etc. This means that we need to define *experimental competences* in a quite broad sense and to fit these into a total set of science competences.

The big challenge is to formulate this holistic understanding of doing experimental science in a way that makes it possible to operationalise it and to find criteria for its performance. Clearly, mainstream hands-on experiments are not doing the job. As Abrahams & Millar (2008) concluded after looking at a wide range of practical lessons:

'… almost all of the student discussion observed by the researcher focused on the practicalities of carrying out the task and, in particular, on who would do what with which piece of equipment…'. (Abrahams & Millar 2008, p. 1957)

What we need is a description of experimental science embedded in the whole process of scientific inquiry including asking the questions, designing the appropriate experiment, collecting the data, analysing the data, interpreting the results and arguing for a model or solution.

Thus, *the second key problem is to establish a valid framework for experimental science*. The understanding of experimental science must be described in a construct capturing the competences we see as central to science practice. The framework must therefore link together the competences involved in doing practical science in accordance with the intended learning outcomes: modelling, empirical work, arguing, communicating, innovating, etc. The more generic competences are especially difficult to conceptualise and new and unknown for teachers to practise.

How to assess experimental science?

It is a central point that a competence is an entity you perform and even if you can describe elements of it, you cannot test these elements of performance separately because it is the ability to link them together into a coherent whole that makes it a competence (instead of independent skills!). The reductionism that has been such a success for science does not work in authentic, human situations. As Swan & Burkhardt (2012) wrote:

'… it was the decision to test the elements of performance separately through short items that undermined the intended performance goals of the 1989 National Curriculum in mathematics in England …'. (Swan & Burkhardt 2012, p. 5)

If a summative assessment is to contribute to improving instruction, the assessment construct must be based on a learning theory relevant for the educational culture and domain that is tested. This could be called *'the construct problem'*: many summative tests have a very simple understanding of student knowledge as something that can be measured along a one-dimensional scale (the accumulation of something) – and not based on research on how students represent knowledge or develop competences in the tested domain. Having designed a valid construct (a model of experimental competence), a theoretically and empirically grounded learning progression must be established, probably with relatively coarse steps, leaving room for teachers' individual and domainspecific filling out. Much work has been and is currently done here (Millar *et al*. 1996; Metzger *et al.* 2013) and much more is needed.

The third key question, then, is *how to monitor and judge student learning processes and level of attainment?* Traditionally this has been done using portfolios and templates and, if introduced properly, the reliability might be sufficiently high. Linked to this is what could be called *'the situation problem'*: assessment of competence needs a rich test environment, and attention must be paid to the social and cultural context, and possibilities of performing processes – like inquiry processes, investigations, etc. Especially including the generic processes causes problems. A recent empirical research project (Nielsen 2015) identified a complex of assessment criteria for assessing students' innovation competence and trialled different examination formats in which students' innovation competence was assessed in parallel with their disciplinary competences and skills.

Another consequence of the situational problem is that the assessment must be performed over a time-span long enough to make it possible for the students to practise tasks reflecting the intended learning outcome in a valid way, whereby students have opportunities to make decisions, select their own method and include many different science processes. Solutions have been tried in many places, typically setting students in a semi-authentic situation and giving them a fixed time-span to go through (most of) the experimental framework. The time-span must be considerably longer than for a traditional examination and might last from four hours to several days.

The project mentioned above designed an examination format in upper secondary school biology that took place over two days. On the first day, each student group worked on a problem concerning blooms of algae in city lakes. The student groups were given a set of materials that briefly explained the problem and contained disciplinary information for them to explore. The students were asked to explore the problem and suggest solutions. During this time, the assessors observed the student groups' work. At the end of the first day each student had to upload a written outline of a presentation. On the second day, each individual student was assessed orally for 15 minutes. Here the student used the talking paper to present the group's findings and suggested solution. Then the assessors could engage with each student in a dialogue about their work. The examination formats and the scoring rubrics provided assessors with enough information to consistently assess the dimensions of innovation competence.

The rationale for establishing a rich and authentic assessment situation can be founded in the situatedness of learning. Just as students' learning is formed by the situation in which they have learned, students' performance simply depends on the situation in which they do the performance (Dolin & Krogh 2010). Together with the wash-back effect of summative assessment on the teaching, these are strong arguments for having a summative assessment that reflect what we find central to experimental science. Tests that only cover a narrow subset of the performance goals will distort the teaching and the learning!

But how should we achieve this in practice? For alignment reasons, the summative assessment must measure the whole range of learning goals as they are ideally realised in the classroom, and it must reflect – and exemplify – the teaching and learning activities in the classrooms. The fourth key question is around how to realise this: *How to combine summative and formative use of assessment?*

We are here at the frontier of assessment research. This very ambitious attempt to combine the formative and summative use of assessment has the big advantage of diminishing the time and money spend on summative assessments – with their poor learning outcomes for the students. Wynne Harlen has advocated for this approach (Harlen 2013). She points to the fact that the experiences that students need in order to develop desired competences also provide opportunities for their progress to be assessed – by the teacher. Teachers can use evidence from ordinary activities supplemented by evidence from specially devised tasks, and typically collected via portfolios. In their article ‘Can teachers’ summative assessments produce dependable results and also enhance classroom learning?’, Black *et al*. (2011) describe some of the conditions that must be met if a dependable system were to be developed.

With the considerable number of competences involved in experimental science, and each competence perhaps having its own learning progression, it will be very time-consuming to build and validate a proper construct. Bennett (2015) argues that ‘…the only way to speed up the process might be to develop and identify theoretically grounded progressions, build them into tests provisionally, use field testing and operational administration to study their validity and impact, and progressively refine the progressions and tests as they are used’.

Integration of IT-based elements is an obvious way forward. The TELS (Technology Enhanced Learning in Science) project is an early example of an IT-based system designed to improve learning and deliver assessments. The project has designed a Web-based Inquiry Science environment (WISE) as an online platform for designing, developing, and implementing science inquiry activities (Linn & Chiu 2011). Such new third generation assessment systems are emerging, trying to serve both institutional (summative) purposes and individual learning (formative) purposes. They typically build a system of complex computer simulations and other performance tasks into a blended learning system. As an example, Bennett (2015) describes how Bennett & Gitomer (2009) designed an integrated system of assessments built into a demonstration program, CBAL (Cognitively Based Assessment of, for, and as Learning). The system would consist of distinct summative and formative assessment components each directed at achieving different purposes.

As Bennett (2015) writes: ‘… a key characteristic of third-generation assessments will be the use of complex simulations and other interactive performance tasks that replicate important features of real environments, allow more natural interaction with computers, and assess new skills in more sophisticated ways’ (p. 387). Students can work with such continuously embedded assessment systems during the year and frequent sampling of evidence of learning is then used formatively and automatically scored and aggregated for accountability purposes. This needs to be organised in a way that is not violating students’ premises for their work and their learning engagement, eg with a clear distinction between formative and summative use of the data and with other ethical considerations, as described by Timmis *et al*. (2016).

Conclusions

The process towards third-generation assessment systems must be based on a political acceptance of the overall purposes of the experimental science and it must design the assessment from a competence model, including learning progressions, grounded in the learning sciences. It must concurrently establish a training and professional development programme and a monitoring system to ensure the validity of the new assessment system.

It is quite encouraging – we know a lot about how to improve assessment of experimental science, but we need a roadmap for realising the vision. As long as accountability is more important than learning, it will be difficult unless we can align accountability and learning purposes. To achieve this, we need to establish a close collaboration between researchers, teachers and policy-makers. A roadmap must thus involve central stakeholders in the process, to enhance their understanding of the central problems and to give shared ownership.

References

Abrahams, I & Millar, R 2008 Does practical work really work? A study of the effectiveness of practical work as a teaching and learning method in school science. *International Journal of Science Education* **30**, 1945–1969.

Bennett, R E 2015 The changing nature of educational assessment. *Review of Research in Education* **39**, pp. 370–407.

Bennett, R E & Gitomer, D H 2009 Transforming K-12 assessment: integrating accountability testing, formative assessment, and professional support. In *Educational assessment in the 21st century* (eds C. Wyatt-Smith & J. Cumming), pp. 43–61. New York, NY: Springer.

Black, P Harrison, C, Hodgen, J, Marshall, B & Serret, N 2011 Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice* **18**, 451–469.

Dolin, J & Krogh, L B 2010 The relevance and consequences of PISA science in a Danish context. *International Journal of Science and Mathematics Education* **8**, 565–592.

Harlen, W 2013 Assessment & inquiry-based science education: issues in policy and practice. Global Network of Science Academies (IAP) Science Education Programme (SEP). (See www.interacademies.net/activities/projects/12250.aspx.)

Linn, M C & Chiu, J L 2011 Combining learning and assessment to improve science education. *Research and Practice in Assessment* **5**, 5–14.

Metzger, S, Gut, C, Hild, P & Tardent, J 2013 Modelling and assessing experimental competence: an interdisciplinary progress model for hands-on assessments. ESERA 2013.

Millar, R, Gott, R, Duggan, S & Lubben, F 1996 Children's performance of investigative tasks in science: a framework for considering progression. In *Progression in learning* (ed. M. Hughes), pp. 82–108. Clevedon: Multilingual Matters.

Nielsen, J A 2015 Assessment of innovation competency: a thematic analysis of upper secondary school teachers' talk. *The Journal of Educational Research* **108**, 318–330.

Osbourne, J 2015 Practical work in science: misunderstood and badly used? *School Science Review* **96**, 16–24.

Swan, M & Burkhardt, H 2012 A designer speaks. *Educational Designer* **2**, 1–41.

Timmis, S, Broadfoot, P, Sutherland, R & Oldfield, A 2016 Rethinking assessment in a digital age: opportunities, challenges and risks. *British Educational Research Journal* **42**, 454–476.

Reflections on current national examination policy changes at 14–16 and 16–18 across the UK: implications for students' experience of the science curriculum and assessment of experimental science

Professor Jannette Elwood, Queen's University Belfast

Introduction
The landscape of current assessment policy across all phases of education in the UK is in a major state of flux. No more so than in the changes we are witnessing to the assessment structures of GCSEs (taken normally by 16 year olds) and A levels (taken normally by 18 year olds) with devolved arrangements for these qualifications across England, Wales and Northern Ireland (Barrance & Elwood 2018). Scotland, too, has seen its qualification landscape change with the implementation of the National Qualifications (National 5 being equivalent to GCSE) (SQA 2016) and reform of the Scottish Highers (taken normally by 17 year olds) introduced and first examined in 2015 (Baird & Gray 2016). Such changes to high-stakes examinations are certain to have major ramifications for young people's future educational and employment opportunities.

This short background paper aims to inform this conference's debate on the assessment of experimental science by placing it in the context of newly reformed qualifications at 16 and 18 and the policy decisions surrounding these. The paper will compare and contrast the assessment structures experienced by young people across three nations of the UK[1] and will aim to highlight significant implications of these differences for their education generally but also the variations that present themselves for young people's experiences of the science curriculum, especially experimental science, and how it is assessed and valued.

National examinations in England, Northern Ireland and Wales 14–16 and 16–18

GCSEs (14–16)
In recent years major differences between the qualification policies of England, Northern Ireland and Wales on the assessment structures of GCSEs have emerged, leading to a degree of variation in the experience of these examinations for teachers and students across the three regions. Specifically, there has emerged an interesting situation where students are studying for the same qualification, labelled 'GCSE', but their experience of these qualifications will be very different depending on where students live and on their teachers' choice of examination specifications and awarding organisations.

Until recently GCSEs were regulated on a three country basis: there was collaboration between Ofqual, the regulatory body for England, the Welsh Government, and the Council for Curriculum, Examinations and Assessment (CCEA), the qualifications regulator for Northern Ireland. However, from 2010, political and educational disagreements regarding the fundamental purposes and quality of GCSEs started to emerge, so much so that by 2013 they became entrenched with the collapse of three country regulation. Since then, major reforms to GCSEs have been implemented in England, many of which have not been adopted in Northern Ireland or in Wales. The consequence is that for the first time there are differences in the ways that students can achieve a GCSE across the three regions. Tables 1 and 2 below outline the main differences in the assessment of GCSEs across English, mathematics (first taught from 2015) and science (first taught from 2016).

[1] This paper focuses on England, Northern Ireland and Wales as they operate under the same (similar) system – GCSEs and A levels – and it is here where most differentiation in experience is occurring.

Table 1. Assessment of new GCSEs by region in English and mathematics									
Region	**English Language**			**English Literature**			**Mathematics**		
	Tiers	**Course structure**	**CA***	**Tiers**	**Course structure**	**CA***	**Tiers**	**Course structure**	**CA***
England	Untiered	Linear	No	Untiered	Linear	No	Two tiers	Linear	No
Northern Ireland	Two tiers	Modular	Yes	Untiered	Modular	Yes	Two tiers	Modular	No
Wales	Untiered	Linear	Yes	Two tiers	Modular	Yes	Three tiers	Linear	No
* Controlled assessment (from Barrance & Elwood 2018).									

Table 2. Assessment of new GCSEs by region for science (biology, chemistry, physics and combined science)				
Region	**Tiers**	**Course Structure**	**CA***	**Practical Experiments as part of exam**
England	Two tiers	Linear	No	No (15% of marks available through examination for demonstrating knowledge of scientific experimentation
Northern Ireland	Two tiers	Modular	Yes	Practical work assessed through CA
Wales	Two tiers	Modular	Yes	Practical work assessed at 10% of GCSE through CA
* Controlled assessment.				

The major changes, outlined above, that are being rolled out with regard to GCSEs are: substantive reviews of syllabus content for core subjects to make GCSEs more rigorous; the removal of controlled assessment in particular subjects; limitations to the number and extent of resits; the introduction of a numerical grading scale at GCSE (9–1) in England, the retention of a letter grading scale (A*–G) in Northern Ireland with the possibility of the introduction of C* grade to provide equivalency for the grade 5 used in England, and the use of both numerical and letter grading scales in Wales.[2] In addition, linear qualifications (assessed solely by examinations at the end of a course) are being re-introduced in England, whereas modular courses will still be available to Northern Ireland and Welsh students. With particular reference to science subjects, there is variation in how experimental science will be assessed and reported across the three jurisdictions with it being assessed through examination questions in England and not reported separately, to it being teacher-assessed and separately reported, for specifications from Northern Ireland and Wales.

A levels (16–18)

The extent of change applying to A level qualifications is no less significant and already the impact of such change has been filtering through to students who commenced these reformed A levels in 2015, with new AS examinations being taken in 2016 and new A levels being awarded in 2017. The

[2] These proposed changes are continually under review and were correct at the time of writing (October 2016).

noteworthy changes to AS and A level examinations are: substantive revisions to syllabus content to make AS and A levels more rigorous, with involvement of higher education institutions in defining subject content; the removal of controlled assessment in particular subjects; the re-introduction of linear qualifications in England with modular A levels still available in Wales and Northern Ireland; and curtailing of the extent of resists. Furthermore, the AS has been de-coupled from the A level in England, thus removing the use of a uniform mark scheme; the AS remains coupled to the A level in Northern Ireland and Wales but with the percentage weight of the AS being reduced to 40%; and gradating scales of A*(A)–E retained for both AS and A levels across the three nations. Table 3 below illustrates these changes as well as how practical science is to be assessed and reported across the different regions.

Table 3. Assessment of new AS and A level examinations by region and inclusion of science practical assessment in biology, chemistry and physics A levels

Region	AS/A level	Course Structure	CA*	Practical Experiments as part of exam
England	Decoupled	Linear	No	No – 15% of marks available through examination for demonstrating knowledge of scientific experimentation Pass/Unclassified grade for practical experiments undertaken in school, teacher assessed and reported on certificate
Northern Ireland	Coupled (AS 40% of A level)	Modular	Yes	Yes – Practical work assessed internally at both AS and A level and contributes 22% of A level overall.
Wales	Coupled (AS 40% of A level)	Modular	Yes	Yes – Practical examination worth 10% of the A level

* Controlled assessment.

Implications for young people of the above policy changes to qualifications

Across the UK, the policy changes applied to GCSEs and A levels will lead to significant variation in the level of control that schools in the different regions will have over how these qualifications have to be administered within their local context, ultimately affecting the Key Stage 4 and A level curricula. These changes may well have long-term impact in the accessibility and relevance of science to young people. In this particular context, differences may emerge in relation to how students will experience science subjects through these reformed qualifications and associated assessment structures depending on the region of the UK in which they reside. Weeden (2011, p. 402) has considered ‘curriculum differentiation’ to embrace not only how the curriculum is organised but also how different schools, over different geographical regions, organise their curricula offer depending on context and local demand, hence differentially impacting on what is available for young people to study. So, for example, the case of AS examinations in science subjects is of interest here. In England in 2016, there was a reduction in AS entries in biology, chemistry and physics by approximately 20% for both boys and girls (JCQ 2016). It is interesting to consider why such a reduction in entries might have arisen. One possibility is, given that AS is no longer attached formally to A level study, there may be fewer opportunities for students to take AS examinations in science subjects as part of a suite of 4+ AS subjects given schools’ decisions around sixth form provision. Consequently, any restriction on choice may well curtail the promotion of diversity of subject choice and the mixing of arts/humanities/science subjects at AS level (UCAS 2015). Another possibility is that some schools may no longer offer AS examinations in a variety of subjects, including science subjects, given that they are standalone qualifications that may well require additional curriculum space and time for study. Keeping track of the degree of impact that these reformed qualifications will bring to bear on the science curriculum will be important for future students' experience of the subjects more generally.

The science subjects chosen by students at 14 and at 16 align very closely with examination specifications. These specifications *de facto* become the curriculum (Madaus 1988) and influence young people's understanding of and exposure to science subjects and how they can become positioned by and within science more generally (Murphy 2008). Within these specifications we see reflected not only the received knowledge domains as to what is deemed appropriate science for young people to know, but also what is considered appropriate ways in which to assess science subjects. The variations in the reformed qualifications with regard to experimental science and how it is to be assessed and integrated (or not) into the final grade is obviously of interest here. Such changes may also differentially affect the perceived value that students will attach to experimental science. During consultations for the new set of qualifications, there was a clear call for the removal of controlled assessment within science specifications to counteract the tendency of teachers to 'teach to', and students to 'learn to', the battery of science practical tests (The Gatsby Foundation and The Wellcome Trust 2013). The main consideration here was the tendency for controlled assessment to promote more of the indirect assessment of practical skills (IAPS) rather than the direct assessment of practical skills (DAPS) (Reiss *et al*. 2012). The outcomes of the devolved arrangements for the assessment of experimental science within GCSE and A level specifications have tried to attend to these limitations and consider more valid ways of assessing practical science while also increasing the amount of DAPS within the specification. The new arrangements (including more teacher assessment across a wider set of experiments) may not necessarily solve all the 'teaching or learning to the practical test' problems but they may create enhanced opportunities for improvement in the teaching and learning of practical science.

However, as with all changes to assessment systems and techniques, other conundrums may present themselves through the new arrangements that may well affect the place of experimental science within the curriculum. It will be interesting to see what the ultimate value will be of reporting achievement in practical sciences through additional statements on certificates rather than through teacher assessed components that are integral to the overall award. This may have particular relevance to those entering higher education and how such statements will be used within these contexts. Furthermore, it will be instructive to consider the variations that will emerge in the ways in which schools will deliver teacher-assessed experimental science; such variations may well lead to instances of Weeden's curriculum differentiation for different groups of students and are also likely to have important consequences for experimental science in the longer term. Evaluating the impact on experimental science of all these changes, across the three regions, will be necessary to understand more fully the fundamental role and place that it will continue to have within the study of science more generally.

References

Barrance, R & Elwood, J 2018 Young people's views on choice and fairness through their experiences of curriculum as examination specifications at GCSE. For a special issue entitled 'Inequalities and the curriculum'. *Oxford Review of Education*. (In preparation.)

Baird, J & Gray, L 2016 The meaning of curriculum-related examination standards in Scotland and England: a home–international comparison. *Oxford Review of Education* **42**, 266–284.

JCQ 2016 *Reformed AS in England 17yrs and under 2015 vs. 2016* (see http://www.jcq.org.uk/examination-results/a-levels/2016/reformed-as-in-england-17yrs-and-under-2015-vs-2016). London: Joint Council for Qualifications.

Madaus, G 1988 The influence of testing on the curriculum. In *Critical issues in curriculum* (ed. E Tanner). *Eighty-seventh yearbook of NSSE,* part 1. Chicago, IL: University of Chicago Press.

Murphy, P 2008 Gender and subject cultures in practice. In Murphy, P & Hall, K (eds) 2008 *Learning and practice: agency and identities*, pp. 161–172. London: Sage/Open University Press.

Reiss, M, Abrahams, I & Sharpe, R 2012 *Improving the assessment of practical work in school science*. London: The Gatsby Foundation. (See http://www.gatsby.org.uk/~/media/Files/Education/Improving%20the%20assessment%20of%20practical%20work%20in%20school%20science.ashx.)

SQA 2016 *National qualifications.* (See http://www.sqa.org.uk/sqa/58062.html.)

The Gatsby Foundation and The Wellcome Trust 2013 *Policy note: Assessment of practical work in science*. (See http://www.gatsby.org.uk/uploads/education/reports/pdf/practical-science-policy-note.pdf.)

UCAS 2015 Unpacking qualification reform: results of the UCAS survey on A level reform. (See https://www.ucas.com/sites/default/files/ucas-unpacking-qualification-reform-jan-15.pdf.)

Weeden, P 2011 Inequality of opportunity and the take-up of GCSE geography. *The Curriculum Journal* **22**, 401–422.

Directions in language assessment

Professor Anthony Green, University of Bedfordshire

Summary

Language assessment has always been closely tied to language education, deriving its definition of what is to be tested from pedagogic models. This has included a move towards the use of more ‘authentic’ tasks for assessment, which mimic real-world language use and involve the integration of language skills. Technology is opening new opportunities for innovative task types. Research strands focus on the manipulation of task features to influence the nature of the language elicited; on how test takers tackle assessment tasks; on the use and interpretation of rating scales and other scoring procedures, including automated scoring, and on the effects of assessment on teacher practices and on society more generally.

What and how are languages assessed?

Defining what is to be assessed is fundamental to any assessment enterprise. Language assessment is grounded in, and continues to be framed by developments in language education. This is as true of assessments of language proficiency that are unrelated to any curriculum (such as tests of the English language abilities of nurses or pilots) as it is to assessments of educational achievement that are embedded in schooling (such as GCSE French).

There has been a global move over the past half century in language education away from a view of languages as abstract systems of rules to be studied as academic subjects towards a view of languages as means of communication to be treated as basic life skills. This has been reflected in assessment in a trend away from decontextualised tests of knowledge of isolated words and points of grammar towards skills-based tests (tests of reading, writing, listening and speaking) that involve demonstrating understanding of texts through comprehension questions, and the ability to produce texts such as essays, reports or short presentations and to interact with an examiner or with other test takers.

Recently, tests have increasingly involved integrating these skills to simulate real-world language use in specified contexts (for example, tests of languages for university study that involve listening to a short lecture and reading a text, then writing an essay based on these). Test development models take account of the congruence between both the conditions for performance and the mental processes engaged by language users under natural and test conditions.

Although the focus for assessment has shifted towards language as communication, the model, even for the more innovative assessments, remains closely tied to pedagogy. Languages are defined in reductive ways for educational purposes. Teachers work to impart a defined, usually prestigious, variety of the language (Parisian French, High-German) and assessment focuses on the extent to which learners are able to approximate to this standard. Effective practical communication, in contrast, always involves strategies of negotiation between speakers. When two speakers do not share a language, gestures, pictures and multiple languages can be used to support understanding. Language assessment has not, as yet, found satisfactory ways of coping with the highly contingent, social nature of interaction or the implication that responsibility for communicative success is shared between participants.

What are some of the major challenges facing languages assessment?
The 'what and how' of language assessments have always been topics for debate and research and challenges of construct definition persist. More complex assessment tasks, intended to better reflect language use in the 'real world', are more difficult to score reliably. Discovering more about how task characteristics (such as time allowed for planning and rehearsal, use of support materials) affect performance and how performance can best be scored continue to be urgent priorities. The increasing complexity of language assessment tasks has inspired some to search for more sophisticated psychometric models (such as cognitive diagnostic assessment) that promise to better capture the multi-dimensional character of language use and to provide insights that can enhance learning. Others, reflecting a more general trend in educational assessment, question established measurement approaches, looking towards more interpretive, but less developed social constructivist paradigms.

Beyond concern for the technical qualities of assessment tools, there are challenges involved in how assessment results are interpreted and used by language educators and policy makers. On the one hand, great trust is placed in language assessments as gate-keeping devices, regulating access to academic and professional opportunities (sometimes straying well beyond the stated purpose of the test in question). The unethical use of language tests for purposes that they cannot fulfil (such as confirming the identity of asylum seekers) has been a particular concern. On the other hand, tests are often regarded by the public with hostility and suspicion: there are accusations that language tests can entrench social inequalities and encourage poor educational practices. Low levels of language proficiency have often been blamed on the use of narrowly conceived tests that encourage teachers and learners to prioritise the learning of wordlists and grammar rules over practical communication.

Such concerns have led to calls to improve the 'assessment literacy' of users of language assessments including improving the public understanding of their uses and limitations. Language education has been pioneering in engaging learners in self-assessment and peer assessment, but teacher training programmes continue to marginalise assessment issues. Language assessment specialists consider that teacher training courses should provide teachers (who are being given increasing responsibility for assessment globally) with better grounding in the design, conduct and scoring of assessments and in communicating with learners, parents and educational agencies about outcomes.

What are the most interesting avenues of languages research that are now being explored?
In language assessment, as in most fields, technology has opened up a wide range of new avenues for both practice and research. Adaptive assessments that can tailor the process to individual candidates promise to overcome traditional security concerns and to generate much richer insights into performance. Automated scoring of writing is well-established and systems for scoring speech are also emerging. Although this remains a distant prospect, assessment could become fully integrated into language tutoring systems. Programmes of research into the qualities of new assessment formats are beginning to show how this can be done.

The use of technology to record performance has provided researchers for the first time with substantial quantities of data from very large numbers of language learners engaged in carrying out defined language use tasks under specified conditions. Resources such as the Cambridge Learner Corpus for English (a 50 million-word collection of exam scripts written by students taking Cambridge English exams around the world) are allowing researchers to find new insights into second language learning, identifying which grammatical features and which vocabulary is associated with learners with different levels of functional proficiency.

Technology such as key-stroke logging, eye-tracking and brain tomography is now being deployed to investigate how language learners approach assessment tasks. This can offer new insights into the mental processes involved in using a second language, allowing comparisons between more and less proficient learners and between processing under test and more naturalistic language use conditions. Alongside this research into test-taking, another interesting strand of research has explored the consequences of test use. Some studies have investigated how teachers and learners adjust their practices when preparing for a test. Others have explored how the use of assessment systems have impacted on educational outcomes and how scores are interpreted and used by employers and policy makers.

Conclusions

Language assessment has a thirty-year history of experimentation with complex and collaborative performance tasks. The issues that these raise, particularly for scoring, remain problematic, but new technology promises practical means of overcoming some of the challenges of administration and interpretation. Concern for the social and educational consequences of assessment use have come to the fore in the new millennium and better public communication about language assessment issues is a priority.

Using digital technology to assess experimental science

Dr Michael K Seery, School of Chemistry, University of Edinburgh

Summary

A modern laboratory education curriculum should embrace digital technologies with assessment protocols that enable students to showcase their skills and competences. With regards to assessment, such a curriculum should:

- incorporate the digital domain for all aspects related to experimental work; preparations, activities, reflections;
- provide a robust and valid assessment framework but with flexibility for individuality;
- emphasise to students the role of documenting evidence in demonstrating skills and competences by means of micro-accreditation, such as digital badges.

This paper summarises how some digital technologies can address the above points.

How can research into the use of digital technology in the assessment of experimental science improve the validity of assessment in the short, medium and long term?

Re-shifting the emphasis of assessment by means of e-assessment

Our use of digital technologies in everyday life has increased substantially in the last two decades. In contrast, laboratory education has remained stubbornly paper-based, with laboratory notebooks at the core of assessment protocols. This emphasis on a *post hoc* report of work done, rather than a consideration of the work itself, means that the value of laboratory work has been distorted in favour of the process of preparing laboratory reports. Experimental work, and the demonstration of experimental skills and competences is of secondary importance.

There are good reasons why emphasis has historically been on the laboratory report instead of laboratory work. Directly assessing experimental work, and indeed any input students have to the planning and completion of experimental work, is subjective. Issues also arise if laboratory work is completed in groups, for either pedagogic or resource reasons. Assigning individual marks is fraught with difficulty.

Digital technologies can provide a basis to address many of the concerns regarding validity that the above issues raise, and provide an opportunity to reposition what is considered to be important in terms of the goals and purpose of experimental science.

The completion of experimental work typically involves:

- **Preparation** – planning and preparing for work and making decisions on experimental approaches to be taken;
- **Action** – learning how to carry out work competently, demonstrating competence in experimental approaches, and accurately recording data and/or observations;
- **Reflection** – drawing conclusions from data, reporting of findings, and evaluation of approaches taken.

Incorporating the digital domain for all aspects of experimental work

Wikis and electronic laboratory notebooks are online document-editing spaces that enable individual contribution to be documented and reviewed. Such platforms have been shown to allow the documentation of students’ thoughts and contributions to work, and as such they provide an excellent basis for recording the entire process (preparation, action, reflection) students engage with while completing experimental work. Preparation can include a description of what equipment will be used and why, or thoughts on the purpose of experiment. Action can be documented by recording experimental work completed with the inclusion of data or observations in a variety of multimedia formats (text, photos, video or audio). Reflection can allow for a richer form of the typical lab report. In practice this means asking students to consider and review their experimental approach, so that the emphasis shifts away from the 'right answer' (an often-cited criticism of students coming through a school laboratory curriculum) and more towards a consideration of the approach taken.

Using traceability as a basis for validity of assessment

Validity is a core concern for a national high-stakes examination. Research to date on wikis has pointed to the advantages offered, including that students' contributions are date stamped, and each individual contribution is logged. Overall contributions to work can be tracked. Rubrics have been effectively used to assess students' laboratory skills, although the compilation of rubrics needs a considerable investment in order to document the desired goals and expectations of any particular curriculum experiment or inquiry so that they can be easily assessed. The value of a more flexible approach to documenting science work using wikis and electronic lab notebooks allows scope for individuality within an overall framework of requirements. However, this is an area that needs considerable and ongoing research.

There is a body of research discussing the use of virtual laboratories for mimicking student experimentation, as these provide for more controlled and hence robust assessment protocols. These should be resisted, as they remove students' exposure to the situational and psychomotor demands that being in the laboratory brings. While virtual laboratories may play some role in summative assessment – for example in decision making – they will likely act as a distraction to the necessary changes required to engaging with and documenting real hands-on work, as they will again shift the focus of experimental science away from actual laboratory work.

Emphasis on experimental science and documenting competences

An advantage of a refocus on documenting of processes means that there is an opportunity for students to showcase their own experimental skills. Digital badges have emerged as a way to accredit these, in what is known as 'micro-accreditation'. Digital badges mimic the idea of Guides and Scouts badges by acknowledging achievements and competences in a particular domain. Examples could include badging students' experimental skills (for example, badges for pipetting, titrating, etc) and higher-level badges (for example, badges where students would need to draw on a range of competences already awarded and apply them to a particular scenario (for example, an overall analysis where students would need to design the approach and draw on their technical competency on pipetting and titrations). This enables students to document their own progress in an ongoing way, and allows them to reflect on any activities needed to complete a full set of badges on offer. This is an exciting area as it offers significant expansion across the curriculum. Mobile learning platforms will make new and interesting ways to develop these approaches.

Conclusions

Changing from paper-based to electronic-based media is not without difficulties. In terms of short, medium, and long-term objectives, an initial focus should begin with promoting the possibilities of documenting scientific work in school through the use of multimedia. This will develop a culture and expertise around the use of technical skills and work towards a medium-term goal of developing a basis for documenting work in an online platform instead of on paper – emphasising the value of documenting evidence of processes. This can be complemented with the development of a suite of digital badges associated with expected experimental techniques and protocols. In the long term, this allows the consideration of assessment of laboratory work via wikis and electronic lab notebooks, using appropriate rubrics, which allow students to showcase genuinely and accurately their competence in experimental science in a much more meaningful and engaging way.

Biographies of the authors

Professor Jens Dolin is Head of Department of Science Education at the University of Copenhagen. His research areas are teaching and learning science (with focus on the development of competences) and organisational change (reform processes, curriculum development, and teacher conceptions). He has participated in and been leader for a number of Danish and international science education research projects (including FP7-projects such as S-TEAM, Mind The Gap about inquiry-based science education and ASSIST-ME) and is member of a number of Danish and international boards and organisations.

After a long career as high school teacher in physics and geography (involving development work and part time teaching at university) he gradually drifted into research and employment in higher education and involvement in educational policy with emphasis on science education.

Jannette Elwood is Professor of Education in the School of Social Sciences, Education and Social Work at Queen's University Belfast. She has worked in the field of educational assessment across research, policy and educational arenas and her research has been funded by national charities, research councils and government departments. She is an executive editor for the journal *Assessment in Education: Principles, Policy and Practice* and has recently been elected as Vice President (President Elect) of the *Association of Educational Assessment Europe*.

Her main research interests are in the social constructions and consequences of tests, examinations and assessment practices. She is particularly interested in socio-cultural approaches to understanding the impact of assessment on students' lives; gender and its interaction with assessment techniques and practices and student voices within assessment policy and research.

Professor Anthony Green is Director of the Centre for Research in English Language Learning and Assessment at the University of Bedfordshire. He is the current President of the International Language Testing Association (ILTA). He has published widely on language assessment and is Associate Editor of *Language Assessment Quarterly,* Executive Editor of *Assessment in Education* as well as serving on the editorial board of *Assessing Writing.* His most recent book Exploring Language Assessment and Testing (Routledge, 2014) provides trainee teachers and others with an introduction to the field. Other recent books include Language Functions Revisited (2012) and IELTS Washback in Context (2007), both published by Cambridge University Press. Professor Green's main research interests concern relationships between language assessment, teaching and learning.

Dr Michael Seery is a Reader in Chemistry Education at the School of Chemistry, University of Edinburgh. His research interests align with two main themes: (1) technology enhanced learning to prepare students for learning and (2) laboratory education. He is a previous winner of the Royal Society of Chemistry (RSC) Higher Education Teaching Award and the (Irish) National Association for Integration of Research, Teaching, and Learning Award. He is Chair of the Editorial Board of *Education in Chemistry*, a magazine published by the RSC for chemistry teachers, and Associate Editor of the journal *Chemistry Education Research and Practice.* He is currently writing a book on laboratory education in chemistry, and maintains a blog on various aspects of education at www.michaelseery.com.