The impact of qualification reform on the practical skills of A level science students

Paper 2: Pre- and post-reform evaluation of science practical skills

May 2018

Ofqual/18/6371

Contents

Authors

This report was written by Stuart Cadwallader from Ofqual's Strategy, Risk and Research directorate.

Acknowledgements

The author is very grateful for the ongoing hard work, dedication and expertise of the staff from the 15 university departments who are participating in this study. The author also gratefully acknowledges the support and expertise of the working group that assisted in the development of the materials and the organisations that have provided advice and encouragement. Finally, the author is grateful for feedback received from members of Ofqual's Research Advisory and Standards Advisory groups.

Executive summary

Reformed A level science qualifications were established for first teaching in September 2015. The post-reform qualifications employ new assessment arrangements whereby the direct assessment of practical skills does not contribute to the primary A level grade but forms a separate 'endorsement', the result of which is reported alongside the primary grade. The intention behind these new arrangements was to facilitate more frequent practical work that is better integrated with course content and is assessed in a valid and manageable way.

The assessment of practical skills is now achieved in 2 ways (see Ofqual, 2015). First, at least 15% of all available marks in written examinations must be allocated to questions that indirectly assess practical skills. Second, each student's practical work must be observed and assessed by their teacher throughout the duration of their studies, during which they must complete a minimum of 12 'hands-on' practical assignments. Students are assessed against criteria which reflect the broad competencies that A level science students are expected to develop and receive a separate grade for their performance (either 'Pass' or 'Not Classified').

Some stakeholders have questioned whether schools may deprioritise practical work as a result of assessment arrangements. Ofqual is therefore conducting a programme of research to evaluate the impact of the reform on the practical skills that are acquired by students. This report describes interim findings from one strand of this programme (see Annex A for details of the other strands).

A quasi-experimental study was designed to compare the practical skills of 2 cohorts: those who had studied pre-reform A levels and those who had studied post-reform A levels. To achieve this, bespoke assessments of practical science skills were administered to new first year undergraduates (prior to any formal teaching) in 15 university departments. Two cohorts of new science undergraduates have been assessed so far, one in autumn 2016 (consisting of students who had taken pre-reform A levels during the preceding summer) and one in Autumn 2017 (consisting of students who had taken post-reform A levels). A bespoke assessment was developed by subject experts for each of the 3 science subjects (biology, chemistry and physics). These assessments require participants to complete a circuit of 5 or 6 separate tasks while under the observation of an assessor, who applies a set of binary assessment criteria to evaluate their performance. To ensure the comparability of outcomes between cohorts, the subject-specific assessments do not change between years.

The results suggest that, overall, the post-reform students outperformed the pre-reform students for biology and there was no statistically significant difference between the 2 cohorts for either chemistry or physics. Self-report questionnaire data that was gathered before the practical skills assessment suggested that post-reform students had undertaken practical work more frequently while studying for their A

levels than pre-reform students.. Overall, the results are therefore encouraging – there is no evidence of a decline in the practical skills of A level students following the reform and some evidence of an increase in practical skills for biology.

However, it is important to be cautious when drawing conclusions for a number of reasons. First, the sample does not necessarily reflect all students who take science A levels. Secondly, assessing practical skills in the laboratory in a reliable way is very challenging because the environment cannot be tightly controlled. Finally, only one cohort of students have completed the reformed A levels and it is necessary to allow the new qualifications to 'bed in'. For this reason, it is important that Ofqual and other stakeholders continue to monitor the impact of the reform over time and try to gain a nuanced understanding of the implications for schools and students. Ofqual will collect data from a third cohort of students in autumn 2018 and report the findings in early 2019.

1 Introduction

Ofqual is undertaking a range of activities to evaluate the recently reformed A level qualifications (Ofqual, 2017b). The introduction of new science A levels, which have been taught in schools since September 2015, brought significant change to the arrangements for assessing practical skills. Ofqual is therefore conducting a programme of research to evaluate the impact of this change (see Annex A). This report describes an ongoing study that is being conducted in collaboration with 15 science departments at 13 universities. The study covers biology, chemistry and physics and seeks to compare the practical skills of students who took the pre-reform science A-levels with those who took the post-reform A-levels. Before discussing the study in detail, this report will describe the new A level science assessment arrangements and the rationale for them. This will be followed by a brief literature review about the definition and assessment of practical skills, setting the scene for a discussion of the research methodology and findings.

1.1 Pre- and Post-reform assessment arrangements

Prior to qualification reform, students taking an A level in science were required to undertake 'Non Examined Assessment' (NEA) to assess their practical skills. This involved students completing a practical activity (or activities) under controlled conditions (usually at their school). The NEA was a component of the overall A level, accounting for 20% of the total marks for the qualification. The exact nature of the assessment varied, depending on the school's chosen exam board and specification. A common approach involved the candidate completing a series of assessed practical activities under controlled conditions. These activities were scheduled and administered by the teacher within a time frame specified by the exam board. Candidates would complete a written script (similar to an exam paper) based on their activity, which was either submitted to the exam board for marking or marked by their teacher and then submitted to the exam board for moderation.

Though this was the most commonly used assessment model, it was not the only one that was available. For example, one chemistry specification required students to conduct and write up their own investigation over a period of several weeks, with the subsequent report providing the main focus of the assessment. Despite such variations, it is reasonable to summarise that the NEA components typically focused on assessing planning and analysis rather than the physical manipulation of scientific apparatus (Abrahams, Reiss, & Sharpe, 2013). It was almost entirely the script (or written report) that was the basis of the assessment.

Following a public consultation in 2013, Ofqual received feedback from teachers and exam boards to suggest that the pre-reform assessment arrangements were not functioning as desired. Initially, 5 issues were identified (Ofqual, 2013, pp. 20–21):

1. The practical skills assessments failed to discriminate effectively between students of different abilities (leading to highly clustered mark distributions and grade boundaries, which were very close to one another).

2. Students' performance in the practical assessments significantly exceeded their performance in the written exams, with a high proportion of students achieving the maximum mark.

3. The approach constrained teaching, overemphasising skills typically required for the assessments at the expense of a broad range of practical skills.

4. Due to logistical challenges, it was not possible for all students to undertake the assessment at the same time. This unavoidable issue was leading to an increase in the incidence of malpractice.

5. The assessment did not produce verifiable evidence of the student's actual *performance* of the practical work, limiting the effectiveness of exam board moderation of teacher marking.

Indeed, Wilson, Wade & Evans (2016) suggest that pre-reform assessment arrangements were 'largely unsatisfactory' in terms of their impact on teaching and learning. Subsequent qualitative research from Ofqual (2017c) suggests that teachers had felt it necessary to teach practical work separately from the core course content because of a need to focus on preparing students for the NEA. This meant that practical lessons were typically taking place towards the end of the course, either in preparation for the NEA activities or as part of the assessment itself. This issue was also noted by Toplis & Allen (2012, p. 7)[1]:

> *"…time constraints, moderation requirements and the dominance of assessment have resulted in practical investigative coursework being restricted to a few tried and tested investigations, divorced from day-to-day science teaching'.*

There was also a sense of unfairness with regard to the outcomes of the NEAs; teachers suggested that it was possible to bend or break rules about the level of support that could be offered to students in a manner that was undetectable to exam boards. This is likely to have been a factor in the skewed mark distributions and high grade boundaries that were observed for many of the pre-reform NEA components (Ofqual, 2013).

At the heart of these concerns is the concept of validity. The assessment arrangements meant that exam boards were being required to prioritise the reliability

[1] Note that this quote was focused on the NEA arrangements at GCSE, though they encapsulate concerns about the similar NEA arrangements that were used for pre-reform A level.

of the assessment over ensuring its validity. The arrangements were not satisfactory in assessing practical skills and were potentially driving unintended and undesirable behaviours in schools. There was a significant concern that the arrangements were encouraging teachers to focus on the nuances of the assessment rather than on the provision of a broader and richer education in practical activities and the development of a range of skills. As part of a public consultation, Ofqual proposed the following:

> Having listened to subject experts in biology, chemistry and physics, we propose that the development of conceptual and theoretical understanding of experimental methods should be assessed in the written exams. We propose that students' abilities to undertake practical work should also be assessed and that these assessments should continue to be marked by teachers, with marking moderated by the exam boards. However, we propose that the outcomes of the practical assessment should be reported on the certificate but not contribute to the overall grade.
>
> (Ofqual, 2013, p. 21)

New assessment arrangements were put in place for the reformed A level science qualifications that were introduced in September 2015. The assessment of practical skills is now achieved in the following 2 ways (see Ofqual, 2015):

1. At least 15% of all available marks in written examinations must be allocated to questions that indirectly[2] assess practical skills. These questions should reflect the experiments, skills and techniques that are to be covered as part of the A level course.

2. Each student's practical work must be observed and assessed by their teacher throughout the duration of their studies, during which they must complete a minimum of 12 'hands-on' practical assignments. Students are assessed against 5 Common Practical Assessment Criteria (the CPAC) which reflect the broad competencies which all A level science students are expected to develop. An overview of the CPAC can be seen in Table 1.

[2] The term 'indirect' is used here to refer to assessment that is based on an examination or report about a practical activity rather than 'direct' observation of the activity as it is performed (Abrahams et al., 2013). Direct and indirect assessment of practical skills is further discussed later in this report, in the literature review.

Table 1. *The Common Practical Assessment Criteria (CPAC)*

No.	CPAC Competency
1	Follows written procedures
2	Applies investigative approaches and methods when using instruments and equipment
3	Safely uses a range of practical equipment and materials
4	Makes and records observations
5	Researches, references and reports

Students are expected to build towards these CPAC competencies and be able to demonstrate them consistently and independently by the end of the course. A typical practical activity does not need to assess all of the competencies at once because the intention is to allow teachers to take a flexible and formative approach to assessment before reaching a final holistic judgement.

Students receive a separate result for the practical 'endorsement' – a ‘Pass’ or a ‘Not Classified’ – when they certificate at A level (it does not contribute to their primary A level grade of A* to E). To achieve the ‘Pass’ grade, the student and their teacher must have accumulated and curated evidence of competency in all 5 of the CPAC. Though the overall approach is formative, the student must have demonstrated themselves to be independently competent in each of the 5 CPAC in order to achieve a ‘Pass’ at the end of their course.

To moderate the assessment, schools and colleges are periodically visited by an exam board ‘monitor’. The role of the monitor is to support teachers in their assessment and ensure sufficient records of activities and achievements are being maintained. The monitor also seeks to ensure that students are being provided with the opportunity to undertake relevant practical work. In the first 2 years, every school offering a science A level was visited by a monitor at least once. In the future, the intention is for schools to be visited at least once every 2 years (though for those schools who are failing to meet the requirements, or are otherwise in need of support, there will be more urgent and frequent scrutiny).

1.2 Challenges and potential issues

The new assessment arrangements are intended to promote practical work in post-16 science. Cambridge Assessment (2016) have suggested that the approach will allow teachers to cover a greater breadth of skills and be more flexible in how they integrate practical work into their teaching. There have also been suggestions that the increased flexibility will allow teachers to take a more holistic approach to planning their courses of study, facilitating a more consistent relationship between practical work and the syllabus (Canning, 2015; Evans & Wade, 2015). Evans & Wade suggest that:

> ‘The potential for the endorsement to genuinely allow students to develop practical skills in a much more meaningful way than has been done in recent assessments, with teachers empowered to use their expert judgement, is very attractive indeed.’
>
> Evans & Wade (2015, p. 66)

However, the changes to the assessment arrangements have sparked debate (Leevers, 2015; Stacey, 2015). Some stakeholders suggested that there may be unintended consequences. The main concern relates to the practical endorsement being graded separately and not contributing to the primary A level grade of A* to E. The Gatsby Foundation (2014) suggested that this element of the reform would ‘risk sending the message to schools and colleges that practical work is of subsidiary importance to textbook learning’ (p. 3). The concern is that practical work will be deprioritised by school and college leaders because it will not be perceived as relevant to the primary A level grades and therefore the measures against which schools are judged. If this were to be the case, it may lead to schools and colleges reducing their investment in the staff, apparatus and facilities required to support high quality practical work (Carter, 2014). Similar concerns relating to the development of practical skills were raised by the Welcome Trust (2014) and the BERG (Biology Education Research Group, 2014).

It is important to note that, arguably, the debate about the assessment at A level is rooted in the wider context of England’s provision for practical science across *all* school years. The Gatsby Foundation (2017) have recently undertaken research in 11 countries (including case study work in 6 of those countries) to explore and exemplify what ‘good’ practical work looks like in secondary education and have identified 10 benchmarks. To ascertain how schools were performing against these benchmarks, approximately 400 schools (10% of those in England) were surveyed between November 2016 and January 2017. Of the institutions surveyed, 36% did not meet any of the 10 benchmarks (though Gatsby point out that they are aspirational and therefore deliberately demanding).

To elaborate, one of the benchmarks was ‘Frequent and varied practical science’ (Gatsby, 2017, p. 12). Based on the international evidence, the report recommended that practical work should take place in at least half of lessons. The survey suggested that, at A level, this was being achieved in only 33% of schools for biology, 55% of schools for chemistry, and 47% of schools for physics (and the proportions were considerably lower at GCSE). Another relevant benchmark was ‘Assessment is fit for purpose’ and emphasises the potentially beneficial role of regular formative assessment. The survey suggested that only 65% of teachers regularly used practical science activities for formative assessment (though qualitative data did suggest that the new assessment arrangements for GCSEs and

A levels may be leading to changes in how formative assessment is used across all school years).

The Wellcome Trust is conducting an ongoing national survey (the 'Science Tracker'), which focuses on the views of GCSE students (Wellcome Trust, 2017a). Around 45% of the students sampled reported doing 'hands-on' practical work at least once a fortnight, with 29% undertaking such practical work less than once a month (or never). The results varied depending on several factors: the course that was being taken, the average academic performance of students at the school (with students in higher performing schools experiencing more frequent practical work), and the socio-economic background of students (with students from more deprived backgrounds experiencing less frequent practical work). Though these results focus on GCSE rather than A level, there is clearly scope for more frequent hands-on practical work in secondary schools and for ensuring that provision is more equitable across schools.

Returning to A level, a related but separate element to concerns regarding the impact of the reform relates to the engagement and motivation of students. It could be argued that students may feel less motivated to complete practical work if they do not perceive it to be strongly relevant to their primary qualification outcome, and might instead focus on learning opportunities which they believe will increase their chances of attaining a good primary grade. If this were to be the case, there may also be a negative impact on students' enjoyment of science and therefore their likelihood of pursuing careers in STEM industries (Wellcome Trust, 2014).

> 'These developments could, through the law of unintended consequences, lead to our students viewing the practical work they undertake as an add-on extra and not essential for scientific understanding.'
>
> (Carter, 2014, p. 13)

This issue was explored during interviews with practicing A level science teachers that were conducted by Ofqual (2017c). The findings suggest that the impact on student motivation may depend on the context of the school, with those working with highly able cohorts sometimes responding less positively about the reform than those working with more mixed ability cohorts. It may be that highly able students will feel less motivated by practical work because they are focused on attaining the highest available primary A level grade (an A* or an A). Since they only need to demonstrate 'competency' in order to attain a Pass grade in the endorsement, these students may believe that their efforts are best directed towards preparation for the examinations rather than excelling beyond the minimum standards required in their practical work. To be clear, such students would still be undertaking the required practical work for the endorsement and to prepare them for the exam questions that

would assess their practical skills 'indirectly'. It is just that their teachers felt they may not prioritise it.

In contrast to this view, many of the other teachers who were interviewed said that the new arrangements will be better at engaging and motivating students in their practical work (Ofqual, 2017c). This perspective centred on 3 primary factors: teachers being able to better integrate practical work into their lessons so that students can better see the relevance to theory; students maintaining their own laboratory books as a record for the endorsement and therefore feeling a sense of ownership for their work; and the assessment taking a more 'formative' approach which allows students to ask questions and feel less anxious about undertaking practical work.

There is further evidence to suggest that the new arrangements are popular among teachers, who believe that the impact on practical skills, among other elements of students' learning, has been positive. The University of Southampton surveyed teacher views with regard to chemistry (Barnes, Laham, & Read, 2017). Though the sample was not large (61 teachers), the respondents were positive about the impact of the reform on practical skills: 13 (21%) said 'hands on practical skills' would be *much stronger,* while 19 (31%) said they would be a *little stronger* (10% said skills would be *a little weaker* or *much weaker*). Similar patterns were observed with regards to students' skills in writing up, interpreting experimental results, and their mathematical understanding. Teachers in this sample also indicated that they felt that the new A levels would be better preparation for study at degree level. A small scale study by AQA (2016) found that teachers were similarly positive about the new assessment arrangements.

Although this early qualitative evidence is largely positive, it is important to remember that the new qualifications have only been in place for 2 years – they have not yet 'bedded in' and it is likely that their effects on teaching and learning will evolve (for better or worse) over time. Sir John Holman's report on 'Good Practical Science' for Gatsby stresses the importance of monitoring the reformed assessment arrangements with regard to their impact on the quality and frequency of practical work (Gatsby, 2017). This is an important point and it will fall to all stakeholders to contribute to this evaluation. The study reported in this document seeks to contribute to such evaluative work. However, before describing the research that was undertaken in detail it is first necessary to contextualise it by discussing some of the key academic literature.

2 Key literature

This section of the report covers research literature that is particularly pertinent to the methodology for the study and is divided into 2 sections. The first section discusses some of the difficulties in defining practical skills and therefore precisely what it is that should be assessed. The second section describes the difference between 'direct' and 'indirect' assessment of practical skills and how these 2 approaches may be used and combined. Note that the literature refers to the assessment of practical skills in general – not necessarily to the specific assessment of science practical skills at A level.

2.1 Defining practical skills

There is no clear consensus regarding the exact skills and knowledge which should be included when defining 'practical skills'. Abrahams et al. (2013, p. 243) suggest that: "Currently, practical skill, as a term, is widely used in school science but is rarely defined with anything like the precision that is typical for 'subject content' knowledge in school science". A useful starting point is to consider the activities that may be involved in practical work. The Science Community Representing Education (SCORE, 2014) suggested that effective practical work comprises the use of technical and manipulative skills, extended investigation (including planning, observing, analysing and evaluating) and the development of conceptual understanding of science. Similarly, an international study conducted by Gatsby (2017, p. 17) identified 5 main purposes for practical science:

1. Teach the principles of scientific inquiry
2. Improve understanding of theory through practical experience
3. Teach specific practical skills, such as measurement and observation, that may be useful for future study or employment
4. To motivate and engage students
5. To develop higher level skills and attributes such as communication, team work and perseverance

The above list appears to be fairly comprehensive, but it is worth noting that even broad purposes are likely to change over time. For example, there is evidence that the last 50 years has seen a significant change in what science teachers view to be the core purposes for conducting practical work and there has been an increase in emphasis on teaching practical work for the purpose of preparing students for assessment (Abrahams & Saglam, 2010). This shift appears to be the result of changes to the curriculum and the corresponding assessment arrangements. An increased emphasis on assessment may have also impacted upon how students perceive practical work and its importance, driving a belief that the goal of practical

work is 'earning marks' as much as it is about learning skills or improving scientific understanding (Keiler & Woolnough, 2002).

Practical work clearly has multiple educational purposes and encompasses a range of skills and knowledge that operate in tandem. This breadth is likely to be why there is a lack of consensus on a definition for 'practical skills' – it is difficult to disentangle the knowledge that is required to undertake a specific practical activity from the technical and procedural skills that are required to perform it. For example, a student may not need to understand Ohm's Law in order to prepare a circuit and take accurate readings from a voltmeter. However, they will need some understanding of how to arrange and connect the components, the technical ability to set up the circuit successfully, and an understanding of how to interpret readings from the voltmeter.

Gott & Duggan (1996, 2002) suggest that the knowledge required to carry out a scientific investigation can be described as 'procedural understanding', which is distinct from conceptual and theoretical understanding. Their model (partially reproduced in Figure 1) suggests that procedural understanding encompasses concepts that are related to the acquisition and validation of evidence – it is based on knowledge that underpins the correct application of practical techniques and the correct use of scientific apparatus. The authors acknowledge that the model is a simplification but, nonetheless, a distinction between 'procedural understanding' and 'substantive understanding' is useful.

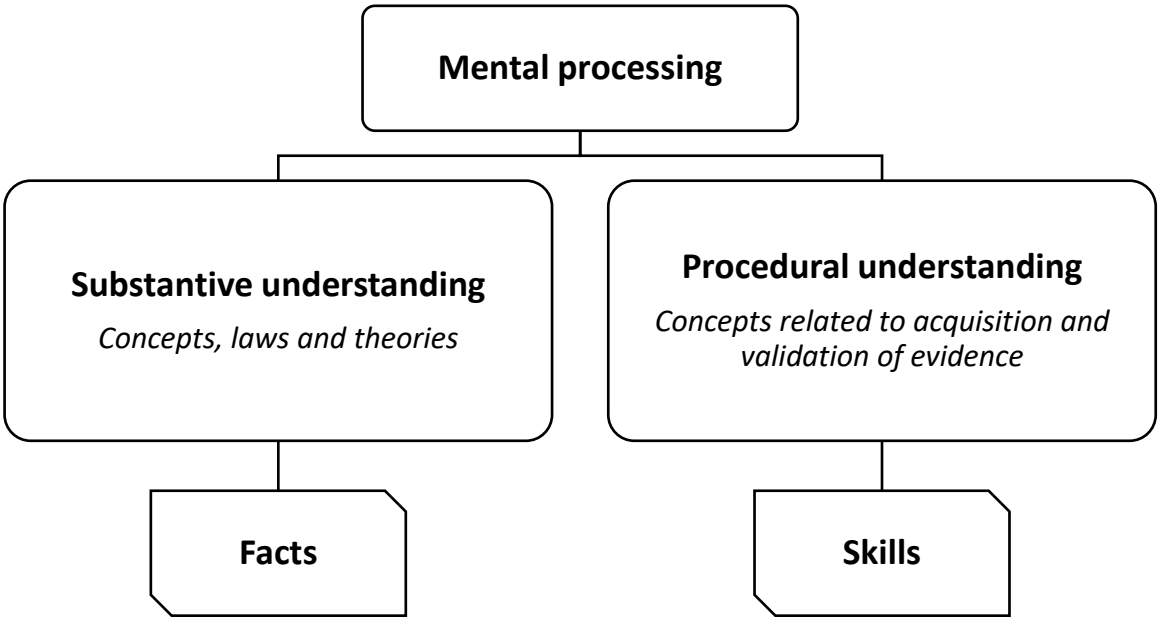

Figure 1. *Gott & Duggan's* (2002, p. 145) *model for problem solving when conducting a scientific investigation*

The element in the bottom right of Figure 1, 'Skills', can be further unpacked, with a distinction between 'Process Skills' and 'Practical Skills'. Process skills are generic in that they are applicable across contexts and practical activties. They include skills relating to 'identifying investigable questions, designing investigations, obtaining

evidence, interpreting evidence in terms of the question addressed in the inquiry, and communicating the investigation process' (Harlen, 1999, p. 129). In many ways, process skills can be described as the 'thinking skills' which stem from an individual's procedural understanding.

On this basis, it is possible to reserve the term 'Practical Skills' to mean only 'those skills the mastery of which increases a student's competence to undertake any type of science learning activity in which they are involved in manipulating and/or observing real objects and materials' (Abrahams et al., 2013, p. 210). Such a definition focuses only on the technical and manipulative skills required for practical work, arguably excluding both deeper conceptual (substantive) understanding and the process skills, which are underpinned by procedural understanding. They could be broadly described as the 'doing skills' which stem from procedural understanding.

For the sake of clarity, this report will use the terminology summarised by Abrahams & Reiss (2015, p. 40), which is in part based on that of Gott & Duggan (2002):

- **Conceptual understanding** – the knowledge base for substantive scientific concepts (eg respiration, atomic structure, thermodynamics) which are underpinned by facts.
- **Process skills** – generic skills such as observation, measurement, sorting/classifying, planning, predicting, experimenting and communication. These are generalisable and transferable between contexts.
- **Practical skills** – performance skills necessary for undertaking non-written tasks (eg performing a titration, reading an oscilloscope). Practical skills are therefore more specific than process skills (eg 'can focus a light microscope at a range of magnifications').

It is important to reiterate that most practical activities will involve both process skills and practical skills, to varying degrees. Many tasks will also rely on some degree of conceptual understanding, particularly if the intention is for them to demonstrate a particular scientific phenomena. Arguably, these elements are inseparable, at least in terms of their assessment (Harlen, 1999). This report does not aim to artificially separate the various elements of knowledge and understanding, just to be as clear as possible about the distinction between them.

Before continuing, it is worth putting all of this in the context of assessment in England. The A level is a 'high stakes' qualification that is based on summative assessment. Assessment outcomes are high stakes for individual students because they partly dictate access to opportunities in higher education and employment. Assessment outcomes are high stakes for schools and colleges because they operate in a market and are held to account for the performance of their students (Acquah, 2013; West, Mattei, & Roberts, 2011).

This high degree of importance, though not unique to England, is certainly more prevalent in England than in many other nations (Mattei, 2012). Hollins & Reiss (2016) conducted a review of the science curricula in high performing nations and suggest that the extent to which practical work is explicitly assessed in schools varies significantly and often relies entirely on school based assessment that is conducted by teachers. They suggest that the assessment arrangements deployed by a nation often reflect the system for school accountability. Where outcomes are 'low stakes', the approach to assessment often focuses less on the reliability of the assessment outcomes and more on giving teachers opportunities to provide structured formative assessment.

On this point, it is arguable that, assuming an unbiased and standardised system, assessment reliability has the potential to be greater in a school-based system because of the large amount of time that teachers spend with their students. Most assessments, written examinations for example, are based on performance over a fairly short period of time (usually 2 to 3 hours). This is just a snap shot of the student, taken at a specific point (albeit at the end of their course). Theoretically, teacher assessment that takes place for the duration of a course is likely to better capture the capabilities of a student over a variety of activities and circumstances (Wiliam, 2001).

The reality is, of course, considerably more complicated. Simulated data suggests that grading accuracy is only modestly better for teacher assessment in comparison to examinations (MacCann & Stanley, 2010). In addition, it is possible that school accountability measures could impact upon the validity of teacher judgements (de Wolf & Janssens, 2007). As we have discussed in the introduction section, this was likely to have been the case with the pre-reform NEA (Ofqual, 2013, 2017c). Those administering the assessment were under pressure to achieve good outcomes (grades) from it, a situation which had the potential to create an uncomfortable conflict of interest.

Finally, Hollins & Reiss (2016) note that assessment arrangements rarely remain static. Japan, Hong Kong, Singapore and Shanghai, which are high performing jurisdictions in the Programme for International Student Assessment (PISA) and have a history of using exam based assessment, are reforming their assessment systems to include more school based assessment of practical work for the explicit purposes of improving pedagogical practices through 'assessment for learning'. What constitutes a good approach to assessing practical work is therefore likely to depend on exactly what is to be assessed *and* to what purpose, 2 variables which are likely to change over time.

2.3 The direct assessment of practical skills (DAPS)

With these issues in mind, let us now consider the methods by which practical skills can be assessed and some of the considerations around these methods. Practical

work has been assessed within A levels since the 1970s and within GCSEs since their introduction in the 1980s. However, what constitutes the best approach to assessment has been a 'vexed question over the last 50 years' (English & Paes, 2015, p. 45). There is a sparsity of literature regarding methodologies for the assessment of practical work, with much of the research focussing instead on pedagogy (Watts & Wilson, 2013) or the impact of assessment on teaching and learning (eg. Buchan & Welford, 1994).

As we have discussed, different assessment models are likely to assess slightly different things, meaning that the validity of the approach will be dependent on what exactly is to be measured. Indeed, Brown & Moore (1994) used factor analysis to compare a practical examination and a teacher assessment which were designed to assess the exact same skills and knowledge in biology. They found that the 2 assessment methods were measuring different aspects of attainment, despite coverage of the same content.

To help untagle this, Abrahams & Reiss (2015) make the distinction between assessment which involves the observation of practical skills in context and assessment which focuses on some product of the practical skill, such as a laboratory report or an exam. The latter of these, which is described as Indirect Assessment of Practical Skills (IAPS), does not assess the performance of practical skills but rather seeks to infer the quality of the performance based on some artefact from it. Direct Assessment of Practical Skills (DAPS), on the other hand, requires the student to demonstrate their level of competence through their performance. Given that this report seeks to evaluate the practical endorsement component of the reformed A level qualification, a component which is based upon teacher observation of practical work as it is undertaken, the focus here will largely be on the DAPS.

For the DAPS, the physical manipulation of objects (the performance) is observed and assessed as it takes place. The relative strengths of weaknesses of DAPS and IAPS are summarised in Table 2 (Reiss, Abrahams, & Sharpe, 2012). The table refers specifically to the assessment of *practical skills* - the strengths and weaknesses of the 2 approaches are likely to differ if they were to also target the assessment of process skills or the theoretical context for the practical work.

Table 2. *Comparison of DAPS and IAPS; reproduced from Reiss, Abrahams & Sharpe* (2012)

	DAPS	IAPS
What is the principle of assessment?	A student's competency at the manipulation of real objects is directly determined as they manifest a particular skill.	A student's competency at the manipulation of real objects is inferred from their data and/or reports of the practical work.
How is the assessment undertaken?	Observations of students as they undertake a piece of practical work.	Marking of student reports written after they undertook a piece of practical work or marking of a subsequently taken written examination.
Advantages	■ High validity ■ Encourages teachers to ensure that students gain expertise at the practical skills that will be assessed	■ More straightforward for those who are undertaking the assessment
Disadvantages	■ More costly ■ Requires teachers or others to be trained to undertake the assessment ■ Has greater moderation requirements	■ Lower validity ■ Less likely to raise students' level of practical skills

As we have alluded to, DAPS was not as strongly represented in the NEA arrangements for pre-reform A levels and GCSE qualifications as one might expect. Abrahams, Reiss & Sharpe (2013) analysed the NEA mark schemes and found that the method for assessment was primarily indirect. Of the total marks available, at most 25% were allocated to the DAPS, and in many NEAs the figure was substantially lower. This is not unusual for science qualifications. IAPS also dominates the assessment of practical work in BTECs, for TIMSS (Harmon et al.,

1997) and for the CREST award[3], with only the International Baccalaureate (IB) having a particular focus on DAPS (Abrahams et al., 2013).

Internationally, a variety of approaches are taken to the formative and summative assessment of practical work, and each country assessing a slightly different menu of practical skills, process skills and conceptual knowledge under the general umbrella of 'practical skills'. Indeed, given the different contexts in which each assessment system operates, it can be difficult to make strong international comparisons. From a broad perspective, China, Singapore and Finland (all nations who perform relatively well in PISA) are particular proponents of DAPS for both formative and summative assessment, with teachers given the responsibility for assessing their students. England (pre-reform), Scotland and Australia tend to emphasise IAPS to a greater extent, relying more on assessment that is provided by awarding organisations (Abrahams et al., 2013).

High stakes summative DAPS can be difficult to achieve as part of a large scale national assessment, at least where a high degree of control is required. Independent testing centres would be one option, but operating these would require substantial financial resource and it would be necessary to overcome significant logistical challenges (eg ensuring that all students are provided with access to a local testing centre). As Roberts & Gott (2006, p. 46) put it: 'Performance assessment requires the pupils to 'perform', and therefore implies the need for observation. This is simply not possible within a mass education context'. The alternative is to conduct DAPS in schools and colleges, but independent and unbiased assessment of individuals in a classroom environment can be challenging due to the social interactions that take place and the limitations imposed by the physical environment (Sund, 2016). It can be argued that such school based assessment '…is coherent and valid, but not especially consistent.' (Prades & Espinar, 2010, p. 457).

There are a variety of forms that DAPS could conceivably take, but it is important to note that many assessment models are not necessarily aligned to either DAPS or IAPS. For example, assessments models which require the candidate to complete an individual scientific investigation (such as CREST) can employ either DAPS, IAPS or a mixture of both. It is dependent on the assessment criteria – a candidate could be awarded a mark for their practical skills entirely on the basis of their report (IAPS) or they could be assessed against a checklist as they undertake their practical work (DAPS).

Another model which can facilitate the DAPS is the Objective Structured Clinical Examination (OSCE), which is widely used for the assessment of medical practitioners and was first conceived in the 1970s (see Harden, Stevenson, Downie, & Wilson, 1975). The exact model can vary, but essentially candidates are observed

[3] Assessment is based on a report from an investigation and a presentation.

as they complete a series of tasks with either real or simulated (standardized) patients and are marked against a set of predefined criteria. OSCEs usually take the form of a circuit or carousel of 'stations', with the candidate having a period of time to complete each station before being moved on to the next (Khan, Ramachandran, Gaunt, & Pushkar, 2013).

A typical OSCE task will focus on a common health issue, requiring the candidate to examine the patient, evaluate the issue, and administer (or discuss) a treatment (Sloan, Donnelly, Schwartz, & Strodel, 1995). However, there is considerable variability in OSCE tasks, with some focussing on practical skills (eg using medical instruments) and others on application of knowledge in context. OSCE stations may also include written tasks and viva style assessments (Khan et al., 2013). In this sense, the OSCE is an umbrella that covers a range of assessment approaches.

Though specific OSCEs have undergone significant research and are understood within the medical literature, the diversity of available OSCEs and stations mean that findings about the overarching methodology are difficult to summarise in a meaningful way. The validity and reliability of the OSCE is highly dependent on the content that is to be assessed, the task design, and factors relating to implementation and administration (Swanson & van der Vleuten, 2013; Turner & Dankoski, 2008). There are some concerns that marking reliability is treated somewhat trivially in the research literature, which can sometimes be somewhat 'vague about psychometrically important details of test administration' (Swanson & van der Vleuten, 2013, p. 23). In reviewing the literature, Turner & Dankoski (2008) found that test-retest reliability coefficients ranged between 0.41 and 0.88, indicating that the reliability of the assessments varied considerably from poor

If the OSCE approach were to be applied to the assessment of practical skills in science it is worth noting that it relies on short tasks that assess specific techniques or areas of practical work. In this regard, the scientific process would be somewhat atomised, as practical work would be distilled into short assessable vignettes. An alternative would be to directly assess practical work as part of a broader project, a scientific investigation. It is important to note that a scientific investigation requires more than just practical skills (as defined in the previous section), with process skills and conceptual understanding, as well as quality of written communication, likely to also be within scope.

A scientific investigation is therefore probably most appropriate for assessing a broader range of knowledge and skills, not just practical skills alone. Indeed, assessments based upon a scientific investigation that have emphasised 'practical skills' over broader 'scientific enquiry' have been criticised in the past (Hodson, 1993) for being too atomistic and therefore lacking validity. Assessment via a scientific investigation would likely blend both IAPS and DAPS, as it would be required to cover a broad range of practical and process skills, including, potentially, skills associated with reporting and presenting research findings.

2.4 Research question

To recap, practical work is now assessed in 2 ways for science A levels. The endorsement seeks to encourage the development of practical skills through ongoing teacher-administered assessment, encompassing practical and process skills that are best assessed through the observation of performance in context (DAPS). Practical work is also assessed indirectly through the examinations (IAPS), largely, though not exclusively, with regard to planning, analysis and interpretation. The aim of this report is to explore the following broad question:

What impact has the reform of A level science qualifications had on the practical skills that are acquired by students?

In doing this, the intention is to investigate whether the new assessment arrangements are having an impact (be that positive or negative) on the practical skills of students. It may be that the reform has had the unintended consequence of de-prioritising practical work in favour of preparation for examinations. If this is the case, one would expect those students who took the reformed A levels to be relatively weak in their performance of practical skills in comparison to those who took the pre-reform A levels. Conversely, the endorsement may be having the intended effect of encouraging frequent and embedded practical work, thus resulting in no difference in the performance of the two groups of students or perhaps a better performance by those students who took the reformed A level.

It is worth clarifying that the purpose of this research is to explore the impact of qualification reform on teaching and learning and not to evaluate the comparability of qualifications from different awarding organisations. This is partly because the awarding organisations have agreed a common approach to the assessment of practical skills for all post-reform specifications. The intention is therefore to evaluate the impact of the assessment arrangements as a whole, at a policy level.

3 Method

3.1 Research design overview

The research design for this study is fairly complex, so it is helpful to summarise the overall approach before discussing the detail. This section of the methodology provides an overview of the design with the materials, participants and procedure elaborated upon in subsequent sections.

This study employed a quasi-experimental design to compare the practical skills of 2 groups of students: those who had recently studied *pre-reform* A level science specifications and those who had recently studied *post-reform* A level science specifications. A bespoke assessment known as a Practical Skills Measure (PSM) was developed for each of the 3 science subjects (biology, chemistry and physics). These 3 PSMs are similar in design and implementation to the Objective Structured Clinical Examination (OSCE). Participants undertake a carousel of 5 or 6 separate 'stations', completing a different practical task at each, while working under the observation of an assessor (see Figure 2).

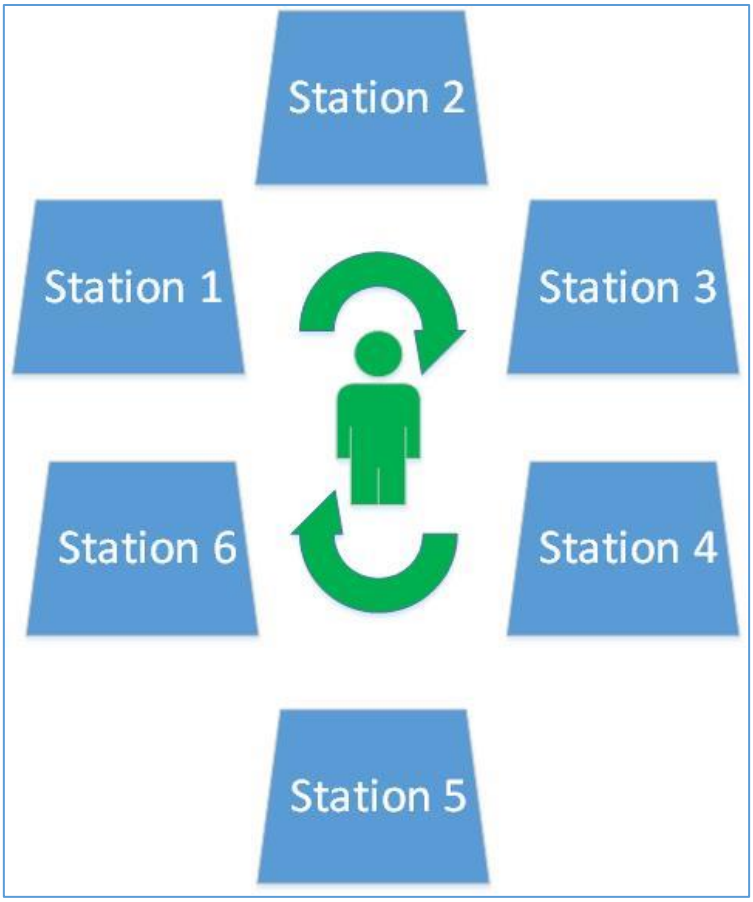

Figure 2. The *PSM carousel*

Participants have 15 minutes at each station of the PSM. For each task the participant is presented with instructions and apparatus and is required to undertake

a specific practical activity without assistance. The tasks were designed to assess techniques that were on the syllabus for pre-reform A levels and are also on the syllabus for their post-reform equivalent. The focus is on process and practical skills that are best assessed by DAPS, rather than planning and analysis skills that may be assessed by IAPS.

The assessor for each station observes the participant as they undertake the task and records whether they meet (or do not meet) each of a series of task specific assessment criteria. These assessment criteria were designed to be binary and as unambiguous in nature as possible (eg it should be easy for the assessor to judge whether the candidate has or has not achieved each of the criterion).

Data collection took place in 15 university departments (across 13 separate universities). Participants were first year undergraduates who had completed science A levels during the preceding summer (more on this later). Data collection took place before participants had received any teaching from their new institution. Participants who had taken pre-reform A level qualifications were assessed from September to November in 2016 (Phase 1), while participants who had taken post-reform A levels were assessed using the same measures from September to November in 2017 (Phase 2). This report compares the performance of these 2 cohorts. A third cohort will be assessed from September to November in 2018 (Phase 3) and reported in 2019.

3.2 Materials

Ofqual commissioned a team of subject experts to develop the PSMs in 2 phases. For the first phase, an initial ‘working group’ was formed that included subject experts from higher education, practicing teachers and academics with specific expertise in the assessment of practical skills. It was this group that suggested the overall design of the PSM and the approach to developing the tasks and assessment criteria. For the second phase, subject specific ‘development teams’ were formed, one for each of the 3 PSMs (this included several members of the working group alongside a number of additional experts that were recommended by members of the working group). These development teams drafted the PSM tasks for their subject to an agreed format and shared these drafts with Ofqual staff for comment.

The development teams were not bound by any particular rules regarding the number of criteria that should be available for each task. However, they were asked to design each criterion to be as unambiguous as possible in order to ensure that assessors would easily be able judge whether or not it had been met. Tasks were designed such that 15 minutes would be a reasonable amount of time in which to complete the activity. The only other requirement was that each task was something that would be on the syllabus for both the pre- and post-reform A level and was best assessed through direct observation of a performance. The tasks were not selected

or engineered to be of any particular level of difficulty, though it was considered important that they allow participants to exhibit a range of performances.

The timing of the study precluded the possibility of piloting the PSM tasks with a representative cohort of participants (eg newly enrolled undergraduates). However, the 3 PSMs were piloted with small groups of second and third year undergraduates. These small-scale pilot studies took place at 2 universities between May and July in 2016. The experience of setting up, running and assessing each task was captured by the university staff (as well as by the Ofqual researcher who was present) and used by the development teams to make refinements to the task instructions and assessment criteria.

For chemistry, it was decided that one of the 6 tasks was not viable due to the duration of the task, which exceeded 15 minutes. This task was therefore dropped from the carousel. For biology, 7 tasks were piloted and 6 selected for the final PSM. The task that was excluded required access to apparatus that may not have been readily available in all universities.

The final tasks for each PSM are summarised in the tables below. For biology (Table 3), some of the assessment criteria can be 'exceeded'. This is where the participant fulfils the requirements of an additional 'Exceed' criterion, which is an extension of one of the main criteria (the number of criteria that can be exceeded are shown in parenthesis in Table 3). Table 5 and Table 6 summarise the tasks and number of assessment criteria for the chemistry and physics PSMs respectively. Details about the criteria themselves are not included in this report because data collection will continue in autumn 2018. Ofqual wishes to minimise the (relatively small) possibility that future participants will access the assessment criteria and gain an advantage.

Table 3. *Biology PSM tasks and number of assessment criteria*

Task	Description	Criteria (Exceed)
1	Making up a standard solution and 10 fold dilution	5 (1)
2	Using a compound high power microscope	4 (2)
3	Determining concentration of unknown from a standard curve	3 (1)
4	Aseptic technique – streaking plates with mock culture	3 (1)
5	Use of an eyepiece graticule	5 (0)
6	Field survey skills	5 (1)

Table 4. *Chemistry PSM tasks and number of assessment criteria*

Task	Description	Criteria
1	Setting up a burette	6
2	Thin Layer Chromatography	6
3	Setting up a reflux and distillation*	7 and 6
4	Making up a standard solution	7
5	Iodine clock (kinetics)	9

Note. Task 3 is divided into 2 parts (the first part involves setting up the reflux and the second setting up the distillation).

Table 5. *Physics PSM tasks and number of assessment criteria*

Task	Description	Criteria
1	Oscilloscope	5
2	Use of micrometre and Vernier caliper	4
3	Measuring resistance	5
4	Preparation of a circuit	7
5	Use of apparatus for timing and a metre ruler	6
6	Using an oscilloscope and a signal generator	5

Before undertaking the PSM, participating students were required to complete a questionnaire which sought to gather information about their qualification outcomes and their attitude to (and experience of) practical work. Participants were asked to provide the subjects that they had taken at A level, their grades and the year that they had completed the qualification. They were also required to provide detail of any other science qualification they had recently acquired. Finally, participants were asked to provide information about how frequently they had undertaken practical work during their A level course and how confident they felt about conducting practical work. The full questionnaire is provided in Annex B. It is important to note that the data gathered from the questionnaire is based on self-report and there is no way to verify that the information that was provided is accurate because the participants are anonymous. This introduces a limitation to the findings.

3.3 Participating universities and students

The decision to work with universities and new undergraduates rather than schools and colleges warrants further explanation. The rationale was in part methodological and in part based on logistical constraints. The A level qualification is primarily for the purpose of assessing “achievement of the knowledge, skills and understanding which will be needed by students planning to progress to undergraduate study at a UK higher education establishment” (Ofqual, 2017a, p. 8). By assessing practical skills in a higher education environment (at the point of entry), the research focuses on the students at whom A levels are primarily aimed – those who intend to continue

studying science at a higher level. This is therefore a self-selecting subgroup of relatively high performing A level science students.

In terms of the logistical constraints, university laboratory environments are more conducive to assessment of this type than school laboratories, which are usually smaller, less well-equipped, and less likely to be available (eg they usually have busier timetables). Conducting the research in schools would also have been more challenging in terms of timing. It would have been important to assess all students at the end of their A level course to ensure that they had received all of their teaching, but it would not have been ethical to risk disrupting preparation for examinations and A level students tend not to come into school during or after the exam period. On balance, it was decided more appropriate and viable to work with universities.

In the first instance, a sample of Higher Education Institutions (HEIs) based in England were approached to participate in the study[4]. University league table data was used to create a sample which was stratified on the basis of the average A level tariff score of new (enrolling) students (see The Complete University Guide, 2017). Initially, only HEIs who offered courses in each of biology (biosciences), chemistry and physics were contacted. This was later expanded to include HEIs that offered degree courses in only one or 2 of the 3 sciences.

This approach elicited only a limited response. This was partially, perhaps, because departmental structures vary both between and within universities, making it difficult to ensure that emails and phone calls are reaching the most appropriate individual. The Royal Societies for Chemistry and Biology and the Institute for Physics kindly supported the project by circulating an invitation to potential participants using their mailing lists of first year undergraduate lecturers and course leaders. This led to a number of universities contacting Ofqual and offering to participate.

Given this mixed approach to recruitment, the final sample should be considered to be largely based on convenience. However, it does include a mixture of Russell Group and non-Russell Group universities and incorporates a range of average A level tariff scores for enrolled students.

Table 7 displays information about the HEIs who participated in both 2016 and 2017. Six participated in the biology PSM, 4 in the chemistry PSM, and 5 for the physics PSM. Note that 2 HEIs were involved in multiple PSMs (eg biology and chemistry / chemistry and physics). For the purposes of this report, these departments are treated as separate HEIs. Precise details of the universities involved have been withheld to maintain their anonymity.

[4] For context, about 190,000 students from the UK were in the first year of a full time degree in a science-based subject in 2016/2017 (HESA, 2017).

Table 6. *Participating University departments and the average A level tariff score for enrolling students*

Subject	HEI	Average A level Tariff score of new undergraduate intake (2017)
Biology (Biological sciences)	B1	350-400
	B2	250-300
	B3	400-450
	B4	350-400
	B5	300-350
	B6	200-250
Chemistry	C1	500-550
	C2	300-350
	C3	550-600
	C4	250-300
Physics (Physics and Astronomy)	P1	300-350
	P2	300-350
	P3	Unknown
	P4	450-500
	P5	500-550

HEIs were paid a fee for their participation in the study. This fee was generally sufficient to cover their costs for materials and for staff time but was not large enough to act as a financial incentive to participation. It constituted a basic fee and a variable fee (per assessor per day), which was dependent on the manner in which the HEI ran the PSM and how many participants they had. The universities who participated therefore did so mainly out of a spirit of collaboration, because they wished to support educational research and the objectives of the study.

Data collection took place at the beginning of the first term, before students had received any teaching from the university. Any additional training prior to assessment (ie training beyond that which the student had received at A level) would be likely to bias the results, so the importance of this was stressed to participating universities. Each university department took a slightly different approach to recruiting students to participate. In broad terms, there were 2 main approaches: either the university would timetable data collection as part of their induction for new students or they would schedule a separate session and invite students to attend. In all cases, participation was voluntary and students took part only if they had read and completed the *informed consent* paperwork. However, it is reasonable to suggest that recruitment to the study was more successful where it was presented to

students as an integrated part of their first year course (albeit one from which they could opt-out).

University departments have a diverse intake of undergraduates – not all of their students complete A level science qualifications. Some of the participants were international students and had taken qualifications aimed at school-leavers in their home country. Other students had come through the English school system but had studied alternative qualifications (eg BTEC) or had taken a gap year (and therefore had not taken A levels in the same year which they had started their degree course). Though some departments invited only students who had completed A levels earlier in the year, most were keen to allow all of their new students to participate in the study, should they wish to do so. Only those students who had completed the relevant A level in the summer prior to data collection are included in the analysis.

3.4 Procedure

The assessment tasks and all supporting materials (instructions, assessment criteria) were identical across all participating institutions. Assessors were usually either university teaching staff or postgraduate students who are also employed as teaching assistants for laboratory classes. The individual in charge of delivering the PSM at each university was responsible for recruiting assessors and ensuring that they had a good understanding of the assessment materials and criteria. In most cases, universities arranged training sessions prior to data collection to ensure that all assessors understood their roles and had a consistent understanding of the assessment criteria.

The basic PSM procedure always involved students completing the questionnaire before embarking on the carousel of stations. Participants had 15 minutes to work on each task while under the observation of an assessor (the end of a period of assessment was usually indicated by a bell or a buzzer that was operated by an assessor with a stop watch). The assessor completes a tick sheet upon which they indicate whether or not the participant has ‘met’ or ‘not met’ each of the assessment criteria. For manageability reasons, assessors often assessed 2 or 3 participants at a time (3 was the maximum permitted). This was deemed acceptable by the PSM developers, who believed that the criteria could be reliably assessed across a small number of participants simultaneously.

It is important to note that university laboratory facilities are not like examination halls in that they cannot be easily standardised. All university departments differ with regard to their facilities, apparatus, availability of staff and timetabling. Though every effort was made to ensure that the same basic model was run at each participating university department, and that the same assessors and technical staff were used for each year, perfect comparability between institutions and years was not a realistic expectation. For example, particular apparatus varied between institutions (eg compound microscopes for the biology PSM), which may have altered the

complexity of the task[5]. To mitigate for this, assessors were told to assist students in locating crucial controls on unfamiliar devices, but were not permitted to explain the function of the control nor to offer any assistance in operating it.

The exact members of staff and postgraduate students involved in setting up, running and assessing the PSM changed unavoidably each year. Though the presence of key staff members was usually stable, assessors often varied between years as new individuals were trained to replace those who had left the university or were otherwise unavailable. Universities made an effort to 'standardise' within their institution to ensure that all assessors interpreted the assessment criteria in the same way and took the same approach to managing the participating students. In the case of one university there was a particularly significant change between years whereby the department moved to a new building. This completely changed the environment in which the PSM was run and, to a lesser extent, some of the apparatus that was used.

In addition, practical constraints meant it was necessary to permit HEIs to run the PSM carousel in slightly different ways. Broadly, there were 3 models for running the carousel:

- **Single laboratory**: This was the 'standard' model for running the PSM and was the most popular. All tasks were set up at stations in a single laboratory and participants rotated between these stations in groups (eg there were multiple assessors and sets of apparatus at each station, allowing multiple participants to be assessed at each). In most cases, each assessor was assigned to a specific station and required to assess between one and 3 participants at a time. In cases where there were fewer participants than there were stations (eg where not every task was running concurrently during an assessment session), assessors would follow participants between tasks.

- **Multiple laboratories:** This model was necessary in 2 scenarios; first, where individual laboratories were too small to accommodate the number of stations or students that were expected, and second, where apparatus required for the different tasks was located in separate laboratories. As with the single laboratory approach, assessors were assigned to specific stations (and therefore rooms) while participants moved between locations. In some cases universities would arrange for guides to escort groups of participants between laboratories (in some cases the laboratories were located in different buildings, which was a challenge for undergraduates who were new to the campus). One university needed to split the tasks up over several days for timetabling reasons. In this case, participating students undertook tasks before (and as part

[5] Ofqual explored the possibility of providing universities with a set of (identical) apparatus but this was not financially viable.

of) particular classes or seminars. Although the administrators reported that the approach was effective, it perhaps constitutes a more significant variation than the other models because it represents a slightly different assessment burden for the participants.

- **Multiple carousels:** Multiple carousels (eg full PSMs incorporating all of the stations) were set up in either separate laboratories or separate spaces within laboratories. These separate carousels were then run simultaneously during an assessment session. Only one set of apparatus was set up at each station for each carousel, meaning only 5 or 6 students were assessed at once. Each carousel had 2 or 3 assessors who would monitor 2 or 3 participants, each of whom were completing different tasks simultaneously. This model was useful where there were larger numbers of participants as it removed the logistical issues associated with moving participants between or across laboratories. It did, however, place a slightly different cognitive demand on the assessors, who were required to consider 2 or 3 sets of assessment criteria rather than just one.

For each student who completed the PSM, the university departments returned an anonymised *Participant Record Sheet* to Ofqual. This comprised the participant's completed questionnaire and the tick sheets for each task in the PSM, showing whether or not the participant had 'met' or 'not met' each of the assessment criteria.

4 Results

What follows is a summary of performance across the tasks for each subject specific PSM and analysis which seeks to explore and explain any differences between the performances of pre- and post-reform cohorts. For each subject, data about the sample and their response to the questionnaire will be presented. Participants overall performance on the PSM is then modelled against whether they are from the pre-reform or the post-reform cohort, their prior attainment at A level, and the university at which they undertook the PSM.

This report will not delve into the nuances associated with each of the practical skills tasks and their underlying assessment criteria. Although this would be of interest to subject experts, it would not be in direct service of the primary research question, which seeks to explore high level differences between the pre-reform and post-reform cohorts.

Before discussing the results, a note on missing data. Where participants have missing data for only one of the tasks in the PSM carousel, they have been included in the analysis. For the purposes of comparison, statistical models in which there was no tolerance for missing data (eg participants were only included in the analysis if they fully completed all of the tasks) were also prepared. There were no substantive differences in the findings.

4.1 Biology

For biology, 6 universities undertook the PSM over the first 2 years (series) of data collection. The 2016 cohort comprised 186 new undergraduates, 140 of whom (75%) had taken A level examinations in summer 2016 (constituting the ‘pre-reform’ cohort). Of the pre-reform cohort, most (138) provided data for at least 5 of the 6 PSM tasks and are therefore included in the analysis. The 2017 cohort was significantly larger due to a change in the approach to recruitment for 4 of the universities that participated[6]. In 2017, 504 new undergraduates participated in the biology PSM, 329 of whom (65%) had sat A levels in summer 2017 (the ‘post-reform’ cohort). Of the post-reform cohort, 298 (91%) provided data for at least 5 of the 6 PSM tasks. Descriptive data about the 2 biology cohorts is displayed in Table 8 and Figure 3. Though a larger number of participants took part in 2017 relative to 2016, the two cohorts are comparable in terms of their achievement at A level (the average biology A level grade of each cohort is very similar and each has a reasonable spread of grades).

[6] These universities were able to integrate the PSM into the first few weeks of the undergraduate course – this was much more successful as a means of recruiting participants than offering the PSM as an extra-curricular activity.

Table 7. *Biology PSM participants (total and eligible) by university department*

	2016			2017		
University	Total	Pre-reform students	Av. Biology A level grade	Total	Post-reform students	Av. Biology A level grade
B1	50	28	B (4.14)	135	73	B (4.37)
B2	48	41	D (2.07)	113	88	C (2.53)
B3	25	21	B (4.00)	79	53	B (3.75)
B4	18	13	B (4.46)	29	26	B (3.96)
B5	14	11	C (3.09)	130	75	C (2.57)
B6	31	26	C (2.92)	18	14	C (3.00)
Overall	**186**	**140**	**C (3.24)**	**504**	**329**	**C (3.28)**

Note: the figure in parenthesis is the mean biology A level grade for the cohort expressed as a number, where A* = 6, A = 5, etc.

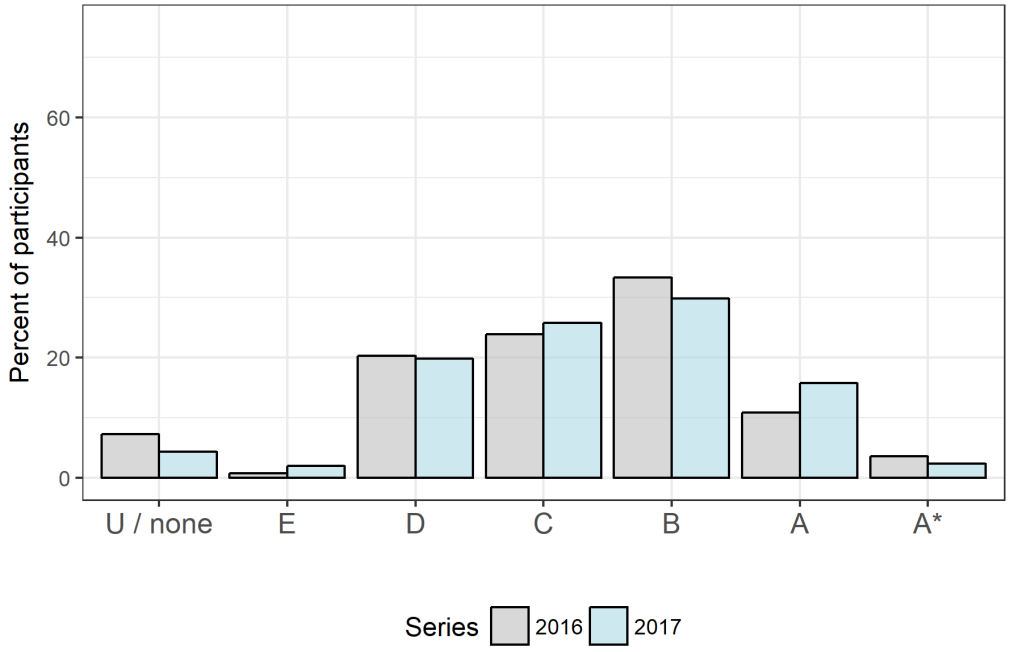

Figure 3. *Percentage of candidates achieving each A level biology grade by cohort*[7]

The plots below display descriptive information about how participants responded to the questionnaire items. Figure 4 relates to each cohort's confidence in undertaking practical work while Figure 5 relates to the self-reported frequency at which each cohort had undertaken practical work during their A levels. If considering the response on a 5 point scale, there was a slight tendency for the post reform cohort (mean = 3.01) to report a higher degree of confidence in their ability to undertake

[7] The relationship between Biology A level performance and PSM performance is presented in a graph in Annex C.

practical work (mean = 2.54, sd = 0.80) than the pre-reform cohort (mean = 2.44, sd = 0.82), but this difference was not statistically significant, *t*(197) = 1.28, p = 0.20, Cohen's d = 0.12. On a 6 point scale, the post-reform group reported doing practical work somewhat more frequently (mean = 3.01, sd = 1.15) than the pre-reform cohort (mean = 2.68, sd = 1.19), and this difference was statistically significant, *t*(231) = 2.61, p = 0.01, Cohen's d = 0.28. However, the manner in which participants responded to these 2 questions did not relate strongly to their performance on the PSM (see Annex C)[8].

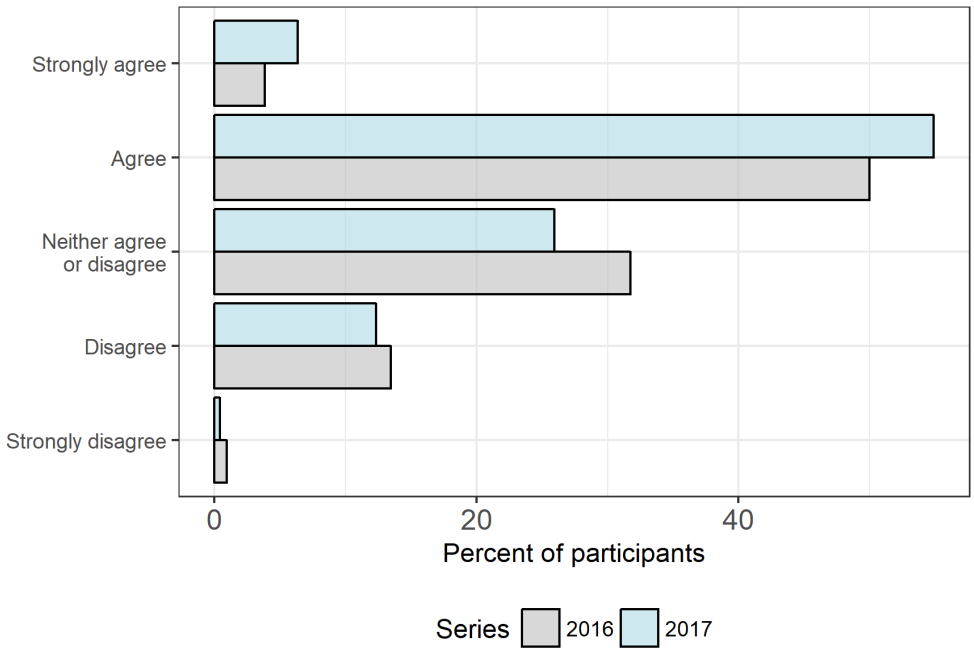

Figure 4. *Participant responses to the question: "To what extent do you agree with the following statement: I feel confident about carrying out practical work" (Biology).*

[8] One of the questions about the frequency of practical work has been excluded from analysis (9. *Please estimate how often you did practical work in your school or college as part of your Biology A-level*). This was because participants tended to replicate their response from the previous frequency question, suggesting that the distinction may have been unclear.

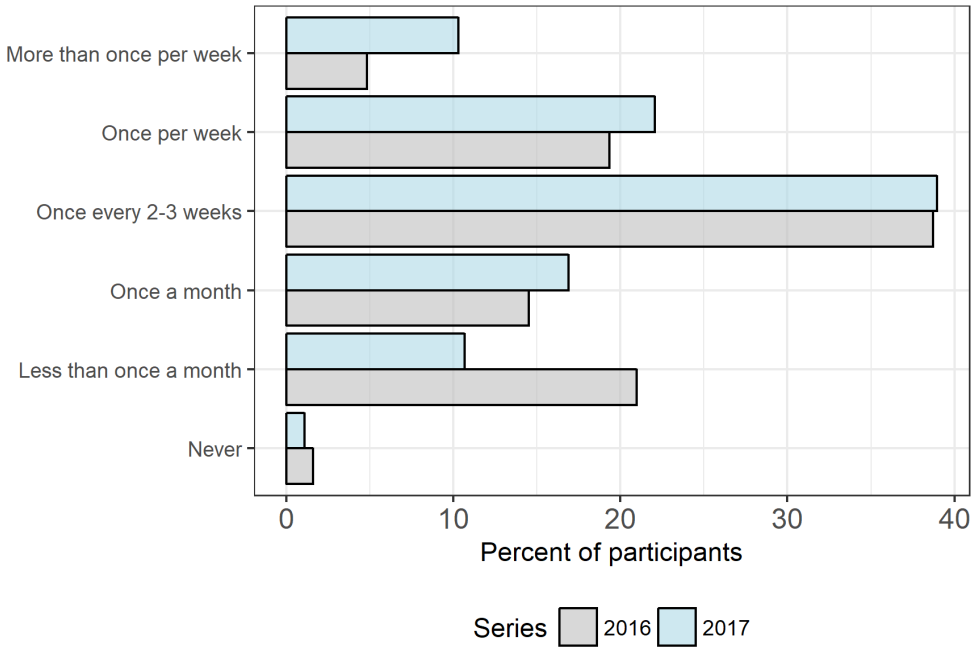

Figure 5. *Participant responses to the question: "Please estimate how often you did practical work in your school or college during your science A levels" (Biology).*

The performance of the pre-reform (2016) and post-reform (2017) cohorts across all universities is summarised in Figure 6. The graph shows the mean percentage of criteria achieved by each cohort for each task but also the mean of these task specific mean values for each cohort (the solid line represents the pre-reform cohort and the dotted line the post-reform cohort). This 'mean percentage of criteria met across tasks' is the primary outcome variable for this report as it represents the average performance across the PSM tasks.

For biology, the participants did not perform particularly strongly on any of the tasks within the PSM. In terms of the mean percentage of criteria met per task, the post-reform cohort (mean = 43.9%, sd = 15.94%) outperformed the pre-reform cohort (mean = 36.2%, sd = 13.70%), $t(307) = 5.17$, $p < .001$, Cohen's $d = 0.52$. This difference can be seen in Figure 6, where the dotted line represents the mean performance for the post-reform (2017) cohort and the solid line the mean performance for the pre-reform (2016) cohort. The task associated with each task number can be found in Table 3 (p.23).

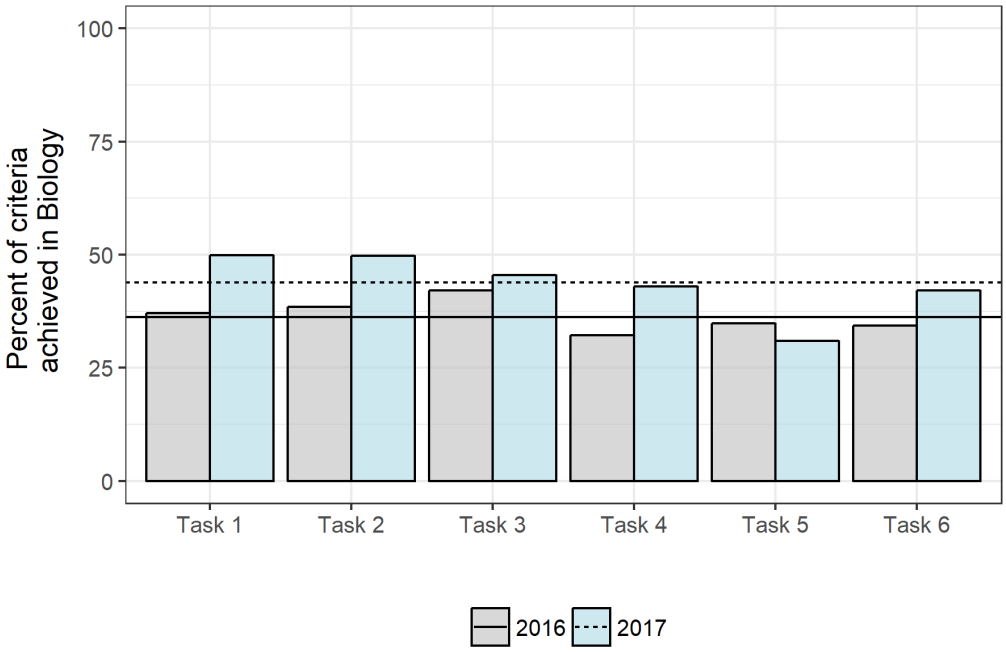

Figure 6. *Percentage of assessment criteria met by task and average percentage criteria met for each cohort.*

A linear regression model was constructed to explore whether overall PSM performance was dictated by particular explanatory variables. As in Figure 6, the outcome variable was 'mean percentage of criteria met across tasks' while the explanatory variables were:

- Whether the participant was from the pre-reform (2016) or post-reform (2017) cohort
- the participant's biology A level grade
- the participant's overall performance at A level (the basis of which is a tariff score calculated by converting all of the candidate's A level grades into numbers and summing them)
- the university at which they were assessed using the PSM

The resulting regression model predicted 38% of the variance and was suitable for predicting the outcome variable (F = 34.14, df = 427, p < .001). The coefficients for the explanatory variables are tabulated in Table 9.

Table 8. *Summary of multiple regression analysis for mean PSM Performance (Biology)*

	B	SE	*t*	Sig.
Constant	21.43	3.71	5.77	<.001
Cohort (2016/2017)	4.73	1.31	3.60	<.001
Biology A level	1.49	0.71	2.09	0.04
A level Tariff score	0.06	0.24	0.25	0.80
University B2	3.58	2.03	1.76	0.08
University B3	1.09	1. 95	0.56	0.58
University B4	16.49	2.36	7.00	<.001
University B5	25.99	2.67	11.47	<.001
University B6	-0.43	2.58	-0.17	0.87
University B1 (reference)	0			

For biology, the results suggest that a small amount of the variance in overall performance can be explained by the cohort, with post-reform participants slightly outperforming pre-reform participants, even when controlling for prior-attainment at A level. However, this difference is not large – all other things be equal, a student from the 2017 cohort achieved about 5% more of the criteria than a student from the 2016 cohort. Prior attainment at A level has less influence in the model than one might expect. The effect of the student's biology A level grade was statistically significant but accounted for a difference of about 1.5% per grade (eg the model predicts that a student who achieved the grade of A* will achieve 4.5% more of the assessment criteria than a student who receives a grade C). The student's general A level performance (across all the subjects which they studied) does not have an impact on PSM performance in the model.

However, there is an important caveat to these findings. The university at which the participant took the PSM predicts their performance more strongly than either their A level grade in biology or whether they took the pre- or post-reform A level. The most notable example of this is university B5, where participants were predicted to achieve an average of approximately 26% more of the assessment criteria than those at university B1 (the reference group for the universities in this model). The mean PSM performance by university can be seen in Figure 7.

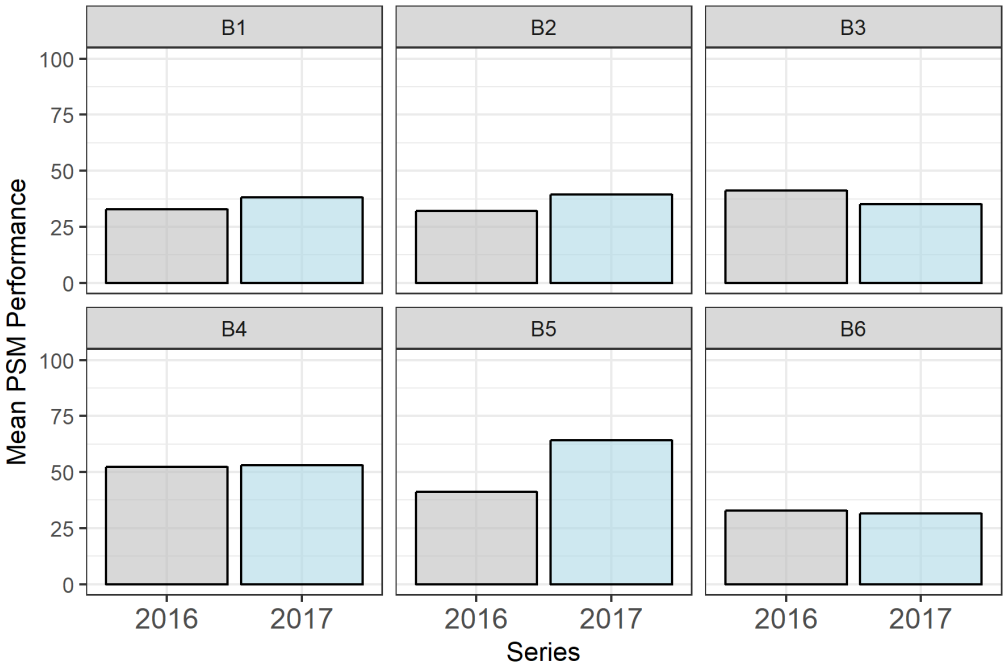

Figure 7. *Mean PSM performance by university (Biology)*

These differences could suggest one of 2 things: a) the application of the binary assessment criteria varied between universities, b) different universities recruited participants with different characteristics and these (unmeasured) characteristics have an impact on the quality of their performance in the PSM. It may also be a combination of these 2 factors. This ambiguity, and the limitation it highlights, are considered further in the Discussion section of this report.

4.2 Chemistry

For chemistry, 4 universities undertook the PSM over the first 2 years of data collection. However, the data for one university was only valid for the post-reform (2017) cohort and so have been excluded from the following analysis[9]. The September/October 2016 cohort comprised 181 new undergraduates, 155 of whom (86%) had sat A level examinations in summer 2016 ('pre-reform' cohort). Of the pre-reform cohort, 109 participants (70%) provided data for at least 4 of the 5 PSM tasks. In 2017, 270 new undergraduates participated in the chemistry PSM, 172 of whom (64%) had sat A levels in summer 2017 ('post-reform' cohort). Of the post-reform cohort, 133 (77%) provided data for at least 4 of the 5 PSM tasks. Descriptive data about these 2 chemistry cohorts is displayed in Table 10 and Figure 8. It is worth noting that most of the participants achieved either grade A* or A in A level chemistry, so there is little variation to explore through statistical analysis.

[9] Though data collection was completed, there was an administrative issue in 2016 which made it impossible to link an individual's performances across the tasks.

Table 9. Chemistry PSM Participants (total and eligible) by university department

	2016			2017		
University	Total	Pre-reform students	Av. Chem. A level grade	Total	Post-reform students	Av. Chem. A level grade
C1	60	48	A (5.44)	129	84	A* (5.71)
C2	0	0	N/A	56	54	B (3.63)
C3	91	77	A* (5.61)	119	66	A* (5.88)
C4	30	30	C (2.93)	22	22	C (2.81)
Overall	**181**	**155**	**A (5.04)**	**270**	**172**	**A (4.98)**

Note: the figure in parenthesis is the mean chemistry A level grade for the cohort expressed as a number, where A* = 6, A = 5, etc. University C2 is excluded from the analysis due to missing data.

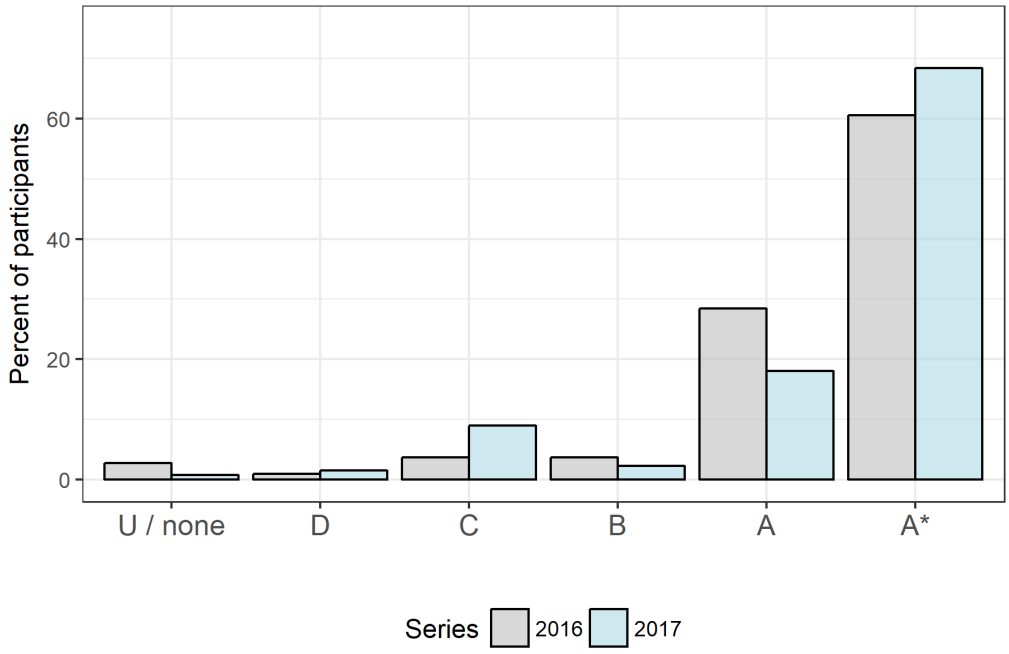

Figure 8. *Percentage of candidates achieving each A level Chemistry grade by cohort*

Figure 9 displays the results from the questionnaire item regarding participant's confidence when undertaking practical work. There was no substantive difference between the post-reform (mean = 2.66, sd = 0.86) and pre-reform cohort (mean = 2.67, sd = 0.80) with regard to their mean level of confidence, $t(184) = -0.08$, $p = 0.94$, Cohen's $d = -0.01$.Figure 10 Figure 10 shows the frequency with which each

cohort reported having undertaken practical work during their A levels[10]. Though a larger proportion of the post-reform cohort reported doing practical work 'More than once a week' in comparison to the pre-reform group, the mean difference in the self-reported frequency of practical work for the pre-reform (mean = 3.41, sd = 0.98) and post-reform (mean = 3.52, sd = 1.08) cohorts was not statistically significant, $t(236) = 0.80$, $p = 0.43$, Cohen's $d = 0.11$.

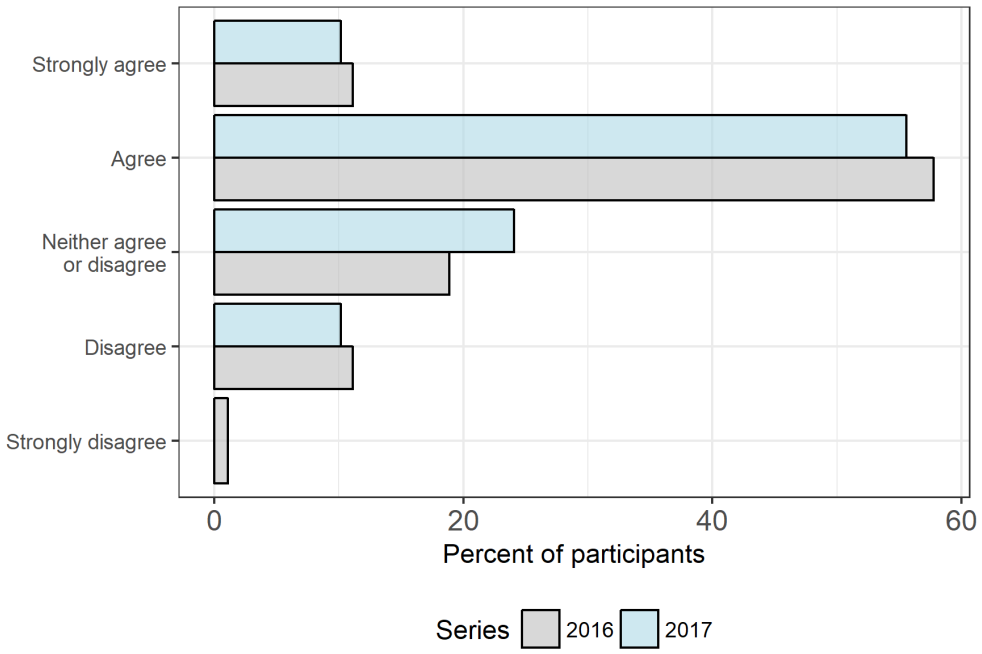

Figure 9. *Participant responses to the question: "To what extent do you agree with the following statement: I feel confident about carrying out practical work" (Chemistry).*

[10] Once again, there was no relationship between these 2 variables and performance on the PSM (see Annex C), so it they were excluded from the statistical models.

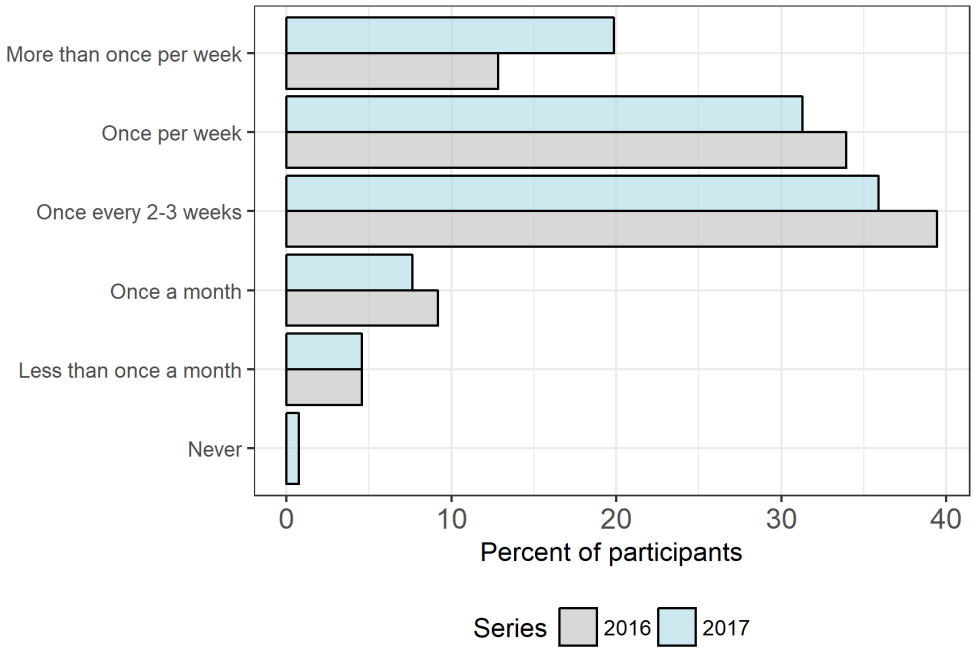

Figure 10. *Participant responses to the question: "Please estimate how often you did practical work in your school or college during your science A-levels" (Chemistry).*

Unlike the biology PSM there is no substantial difference between the overall performance of the pre- and post-reform cohorts (Figure 11). The mean percentage of criteria met (across all tasks) by the post-reform cohort was 67.28% (sd = 14.16%), while the mean performance for the pre-reform cohort was 66.23% (sd = 10.93%). This represents a very small difference which is not statistically significant ($t(239) = 0.65$, $p = 0.52$, Cohen's $d = 0.08$). Despite the similarity in the mean performance of the 2 cohorts, there were notable variations at the task level. The task associated with each task number can be found in Table 5 (p.24).

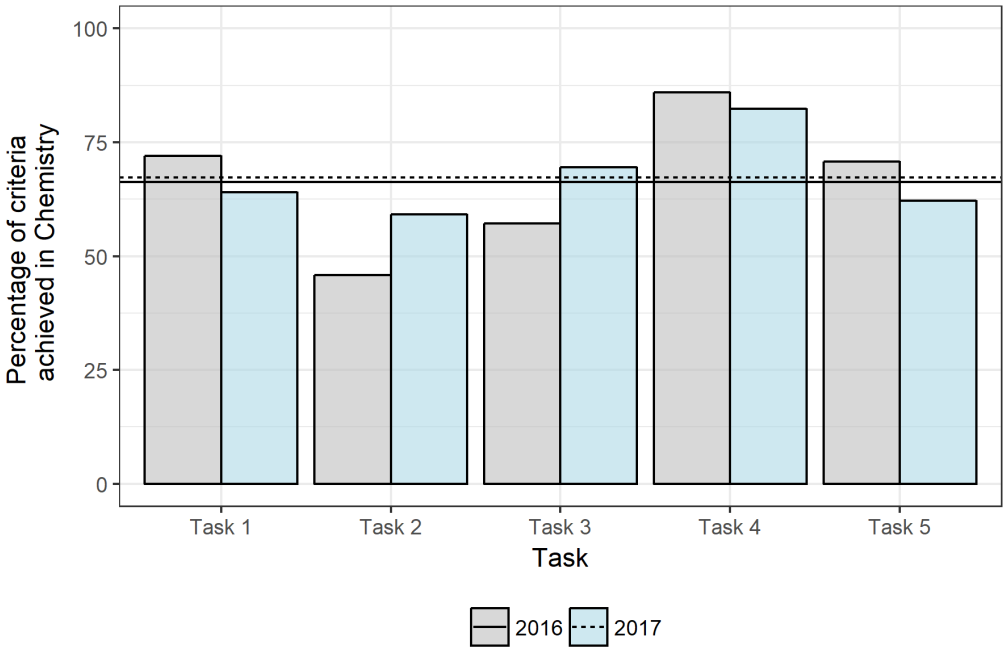

Figure 11. *Percentage of assessment criteria met by task and average percentage criteria met for each cohort.*

The chemistry data were modelled in the same manner as the data for biology. The regression model predicted only 9% of the variance (F = 5.55, df = 236, p < .001). The coefficients for the explanatory variables are tabulated in Table 11. Once again, there were differences in the patterns of performances exhibited by each of the 3 participating universities (Figure 12).

Table 10. *Summary of multiple regression analysis for mean PSM Performance (Chemistry)*

	B	SE	*t*	Sig.
Constant	54.37	5.88	9.25	<.001
Cohort (2016/2017)	0.77	1.62	0.47	0.64
Chemistry A level	0.72	1.14	0.63	0.53
A level Tariff score	0.21	0.21	0.98	0.33
University C3	6.53	1.72	3.80	<.001
University C4	1.65	3.81	0.43	0.67
University C1 (reference)	0			

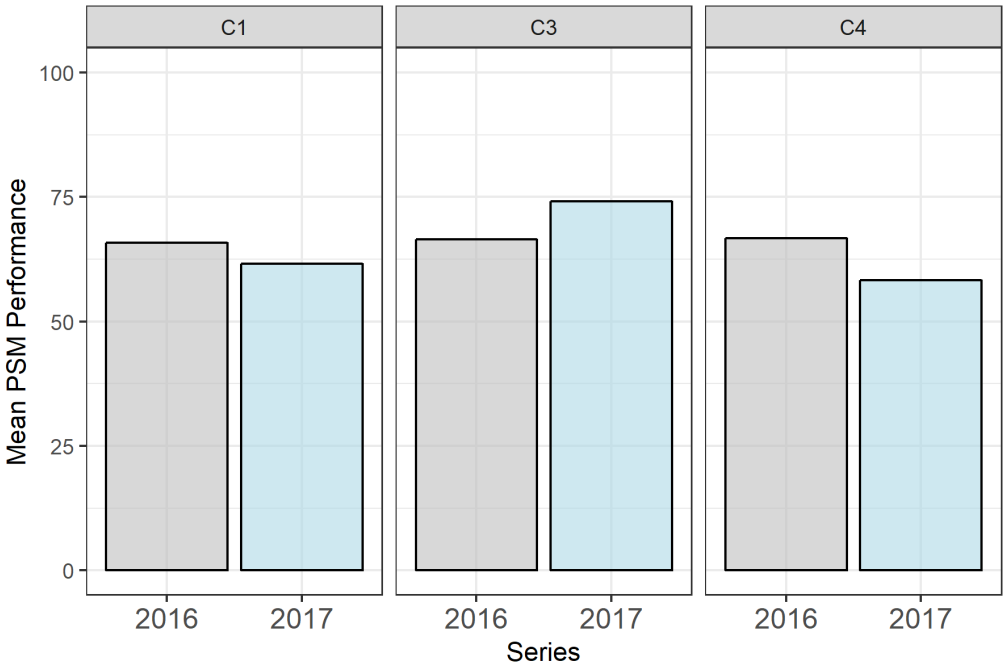

Figure 12. *Mean PSM performance by university (Chemistry)*

Unlike biology, there was no statistically significant difference between the performances of the pre- and post-reform cohorts who took the chemistry PSM. Neither the participant's A level grade in chemistry nor their overall performance across their A levels (their tariff score) predicted their performance on the PSM, though this is likely due to ceiling effects in the attainment data (eg most participants in both cohorts received either a grade A* or A in chemistry). The only variable that appeared to predict performance was the university at which the PSM took place, with, for example, students from University C3 predicted to meet an average of 6.5% more of the criteria per task than those from University C1 (the reference group).

4.3 Physics

For physics, 5 universities undertook the PSM over the first 2 years (series) of data collection. The September/October 2016 cohort comprised 344 new undergraduates, 293 of whom (85%) had sat A level examinations in summer 2016 (the 'pre-reform' cohort). In 2017, 275 new undergraduates participated in the biology PSM, 225 of whom (82%) had sat A levels in summer 2017 (the 'post-reform' cohort). Descriptive data about the 2 physics cohorts is displayed in Table 12 and Figure 13. In terms of the A level grade for physics, participants from both cohorts tended towards the higher grades (particularly grade A), though there was little difference in the average grade achieved by each cohort.

Table 11. *Physics PSM Participants (total and eligible) by university department*

	2016			2017		
University	Total	Pre-reform students	Av. Phys. A level grade	Total	Post-reform students	Av. Phys. A level grade
P1	8	6	D (1.83)	17	9	D (1.78)
P2	43	36	C (3.00)	13	13	C (3.00)
P3	8	8	C (2.63)	5	5	C (3.40)
P4	161	134	B (4.47)	118	98	A (4.60)
P5	124	109	A (5.16)	122	100	A (5.10)
Overall	**344**	**293**	**B (4.44)**	**275**	**225**	**A (4.59)**

Note: the figure in parenthesis is the mean physics A level grade for the cohort expressed as a number, where A* = 6, A = 5, etc.

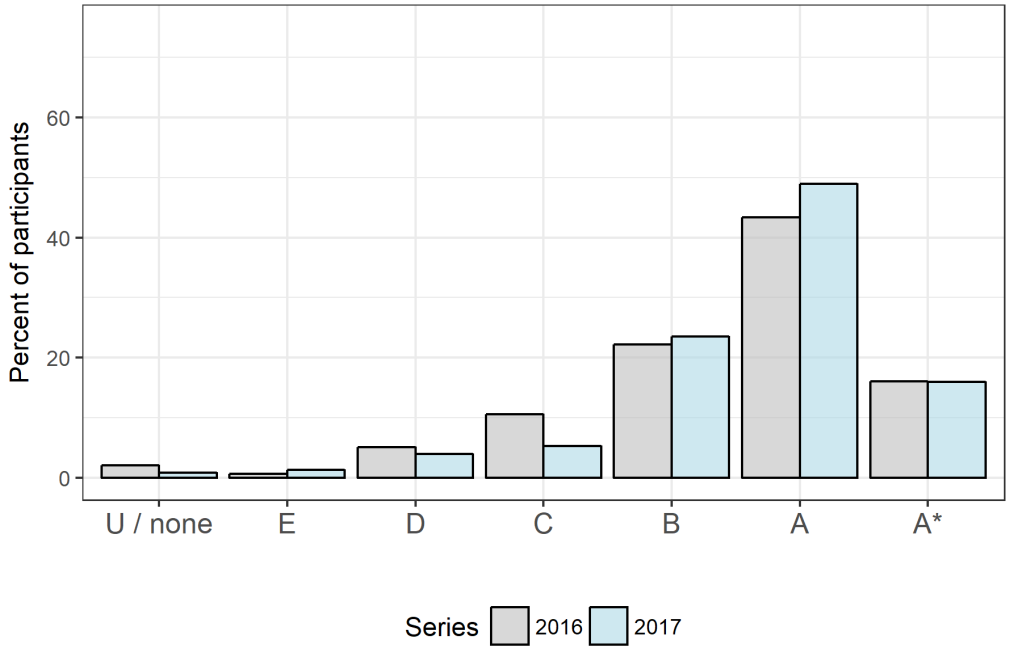

Figure 13. *Percentage of candidates achieving each A level physics grade by cohort*

Figure 14 shows the response to the question about confidence in undertaking practical work. The post reform cohort (mean = 2.46, sd = 0.94) expressed a higher degree of confidence than the pre-reform cohort (mean = 2.18, sd = 1.04), $t(416) = 2.92$, $p = 0.004$, Cohen's d = 0.28. The post-reform cohort (mean = 3.26, sd = 0.88) also reported doing more practical work during their A levels than the pre-reform cohort (mean = 2.80, sd = 1.16), $t(503) = 4.98$, $p < .001$, Cohen's d = 0.45 (see Figure 15).

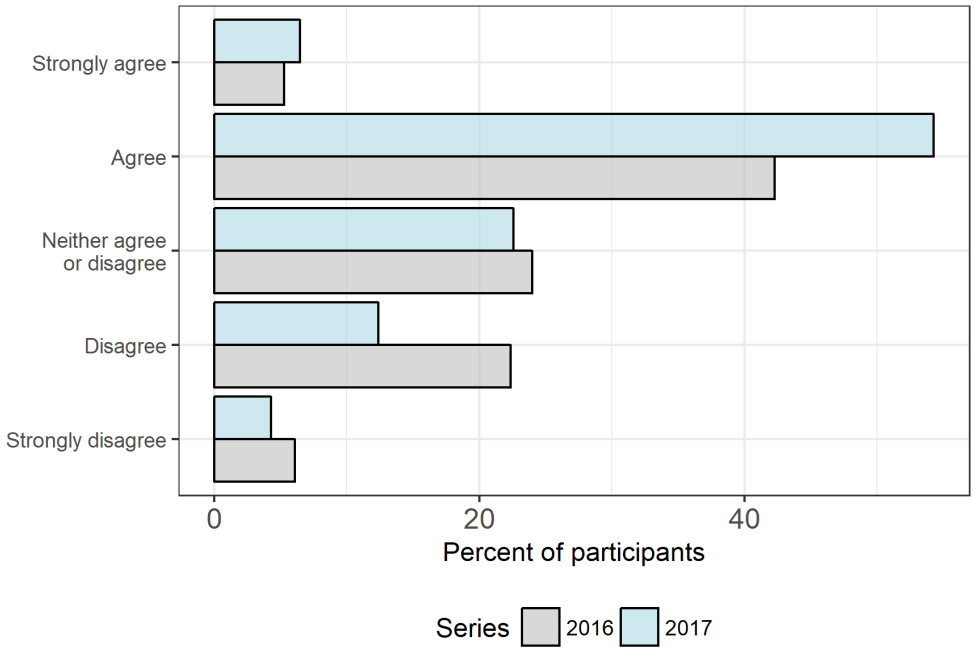

Figure 14. *Participant responses to the question: "To what extent do you agree with the following statement: I feel confident about carrying out practical work" (Physics).*

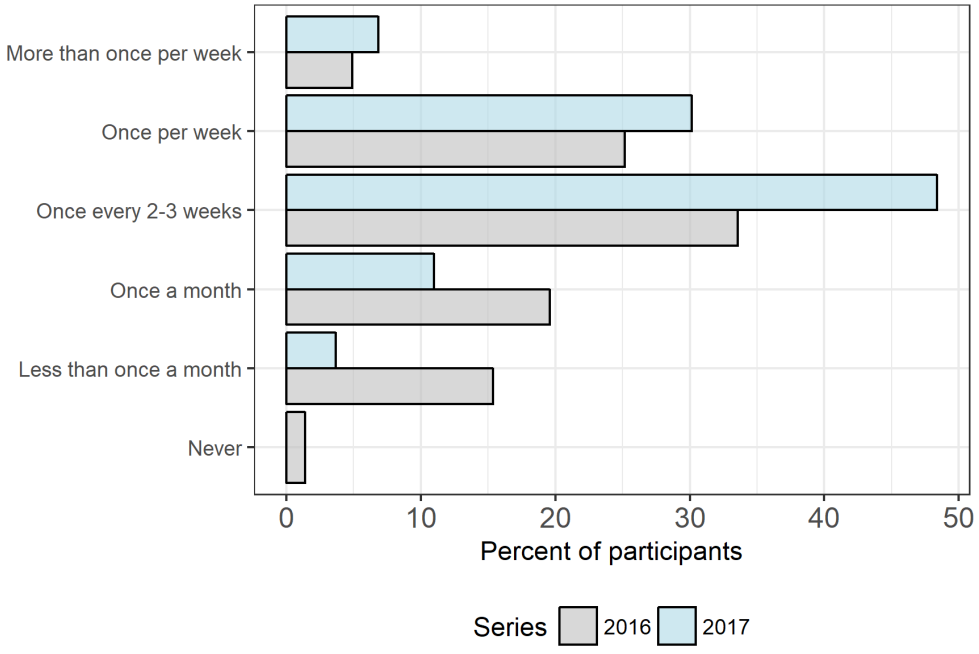

Figure 15. *Participant responses to the question: "Please estimate how often you did practical work in your school or college during your science A levels" (Physics).*

As shown in Figure 16, there was no substantial difference in the PSM performance of the post-reform cohort (mean = 61.13%, sd = 14.44%) and the pre-reform cohort (mean = 59.99%, sd = 13.16%), t(458) = 0.93, p = 0.35, Cohen's d = 0.08. The task associated with each task number can be found in Table 6 (p.24).

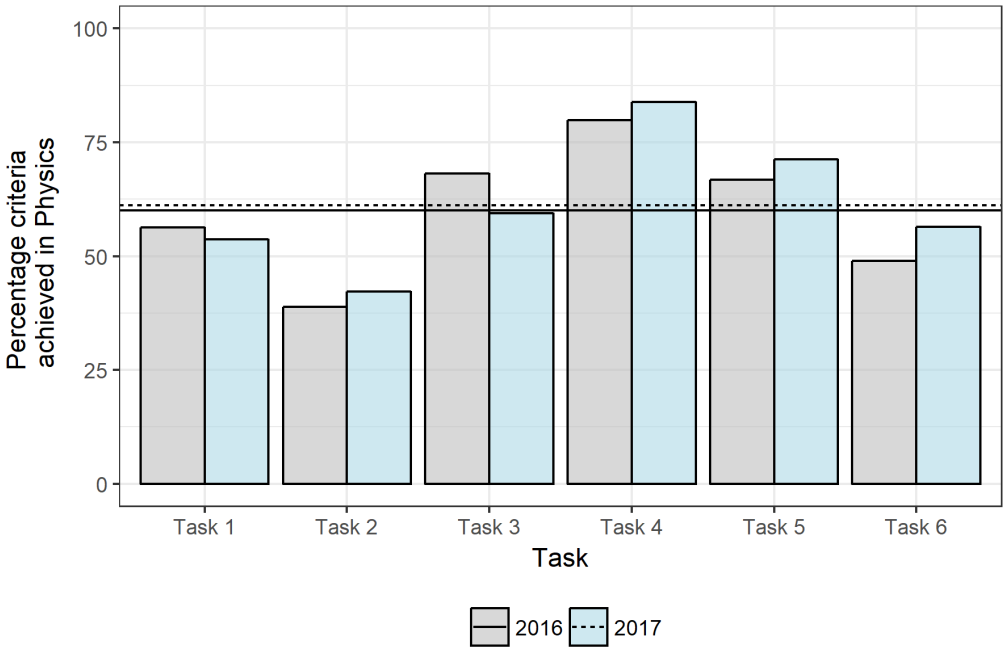

Figure 16. *Percentage of assessment criteria met by task and average percentage criteria met for each cohort.*

As with chemistry and biology, the data was modelled using regression analysis. The regression model predicted only 4.8% of the variance ($F = 4.72$, df = 510, $p < .001$). The coefficients for the explanatory variables are tabulated in Table 13 and the differences between the universities are displayed in Figure 17. None of the explanatory variables predicted PSM performance to a statistically significant degree, with the exception of the participant's A level tariff score. Like chemistry, the pre- and post-reform groups therefore do not appear to have performed differently.

Table 12. *Summary of multiple regression analysis for mean PSM Performance (Physics)*

	B	SE	*t*	Sig.
Constant	50.28	4.02	12.51	<.001
Cohort (2016/2017)	0.69	1.20	0.57	0.57
Physics A level	0.43	0.81	0.53	0.59
A level Tariff score	0.55	0.20	2.74	.001
University P1	-1.57	4.29	-0.37	0.72
University P2	1.32	2.63	0.50	0.62
University P3	-3.55	4.12	-0.86	0.39
University P4	-2.32	1.36	-1.71	0.09
University P5 (reference)	0			

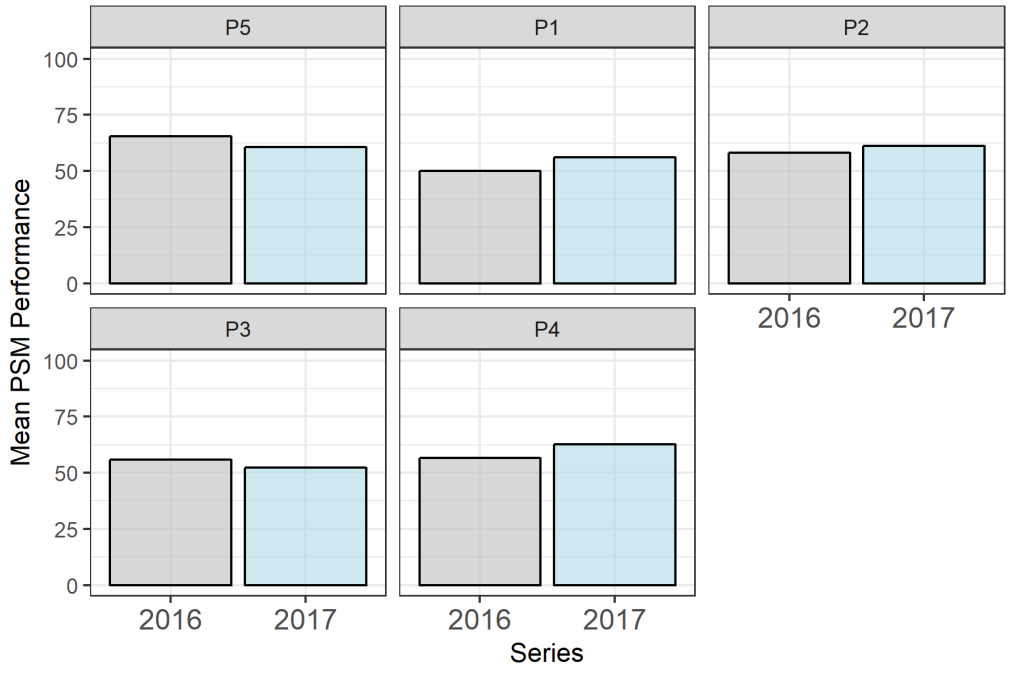

Figure 17. *Mean PSM performance by university (Physics)*

5 Discussion

5.1 Performance by subject and task

Overall, the results show no difference in the performances of pre-reform and post-reform cohorts for chemistry and physics. However, in biology the post-reform cohort outperformed the pre-reform cohort. It is unclear why there was a difference in biology but not chemistry or physics. There are, of course, a myriad of differences between the subjects and the practical activities that take place within them. It seems reasonable to suggest that the reform will interact with each subject in a different way, with a different magnitude of impact. It is not just the assessment arrangements that have changed, but also much of the course content and structure. Teachers and students are adapting to the new qualifications so the impact is likely to change, for better or worse, over time. It is encouraging that the post-reform cohort of students reported undertaking practical work more frequently during there A levels than the pre-reform cohort, suggesting that the reformed qualifications are leading to change.

It is notable that the biology cohorts were more diverse in terms of the A level grades that the students achieved. One theory that could be explored further, and is supported by qualitative work that has been previously conducted (Ofqual, 2017c), is that the reform is having a differential impact depending on the typical academic achievement of the school and/or student. High performing schools may have already been undertaking practical work more frequently (Wellcome Trust, 2017b), meaning that the reform has had less impact on practical skills (though it does not appear to have had a negative impact). High achieving students may be less inclined to focus on the performance of practical work beyond gaining a pass in the endorsement, instead focussing on their examination (including items which assess practical work indirectly). On the other hand, the requirements of the endorsement may be encouraging more practical work in schools where it was employed less regularly, increasing the practical skills of students who tend to achieve grades B-C rather than A*-A. This is speculation, but it is worth further investigation.

The average overall performance of participants was not particularly high in either series for any of the subjects. For biology, the mean percentage of criteria met was around 40%, while it was around 67% for chemistry and 61% for physics. Though this level of performance may seem quite low, it is important not to over interpret these figures. There are 2 reasons for this. First, even though the PSMs reflect practical skills which are expected to be taught at A level, the assessment takes place several months after the participants have gained their qualifications. This means that the PSM is being used as an assessment of how well participants have retained their skills between the completion of A level and the start of university. Realistically, one might expect some attenuation in the skills acquired at A level during this fallow period of time. Second, and more importantly, the PSMs were not calibrated to be of a particular difficulty or to equate to each other. There is no

'benchmark' for a pass. Instead, the PSMs are designed to be an instrument for comparing the skills of the pre- and post-reform cohorts in a consistent way that is free of systematic bias. The sensitivity that the PSM has to differences in performance is more important than the level of performance against the criteria.

Moving on, this study has largely focused on overall performance in each PSM, but there are also notable differences in how the cohorts performed on a task by task basis. One possible explanation for these differences is that the pre-reform students may have focused on particular practical activities if they were prominent in their assessment (the ISA/EMPA components of the pre-reform A level). This could at least partially explain some of the differences at a task level but, without details of the exam board that each student were assessed by (and therefore the details of the NEA they undertook), this is speculation. Once the third and final cohort of students has taken the PSM, it may be worth exploring inter-task differences in detail with the help of subject specific expertise.

Finally, the PSM assesses the performance of practical skills in a specific and somewhat atomised manner. The PSM is focused on process and practical skills that are relevant to performance and are most validly assessed by a DAPS approach. What the PSM does not tell us is whether pre-reform and post-reform cohorts have had different experiences with regard to conducting full scientific investigations, which involve a more complex cocktail of skills and knowledge. The intention behind the reform is that teachers are empowered to conduct scientific investigations with their students and the associated skills and knowledge are assessed by both the endorsement (against the CPAC) and via IAPS in the examinations.

5.2 Limitations of the findings

There are 2 primary limitations to this study. The first of these relates to the sample, specifically whether it adequately represents the population that we are interested in (essentially this is all students who study an A level in science). The second issue relates to the reliability (and therefore validity) of the PSM as an assessment instrument.

Let us begin with the sample. Based on data from the Joint Council for Qualifications (JCQ, 2017), roughly 62,000 students take an A level in biology per year. This means that the sample of 140 that was achieved for the biology PSM in 2016 represents about 0.2% of this population, while in 2017 about 0.5% of the population were assessed. Around 52,000 per year take an A level in chemistry, meaning that the 2016 and 2017 samples each represent about 0.3% of the population. For physics (where the annual entry is about 36,000) about 0.8% of the population were sampled in 2016 and 0.6% in 2016.

Though these samples are, on the surface, quite small, they are sufficiently large to confidently detect moderately sized differences between the cohorts (if we assume

the 2 samples are random). For example, power analysis[11] for the chemistry PSM (which has the smallest sample sizes) suggests that the minimum effect size that can be detected, given a power of 0.95 and $p < .05$, is 0.4. This means that the sample is sufficient to alert us to medium or large differences between the performances of cohorts, just not smaller ones. In other words, the chemistry sample is sufficient to rule out a large or medium impact of the reform on the practical skills which are assessed in the PSM. With regard to biology, the effect size that was observed for the sample ($d = 0.52$) suggests a power of 0.99, allowing us to be 99% confident that there is a genuine difference in the skills of the pre- and post-reform student populations.

Let us return to the assumption that these samples are random (and therefore represent the overall populations of interest). Given the voluntary nature of participation, both in terms of the recruitment of universities and in terms of recruiting participants within those universities, it was not possible to achieve cohort samples of equal size. This was particularly true for biology, where the 2017 (post-reform) cohort is twice the size of that from 2016 (the pre-reform cohort). Despite this, it is worth noting that there was no difference in the average biology A level grade of the 2 cohorts, suggesting that the 2 groups were academically comparable.

For chemistry and, to a lesser extent, physics, the sample is biased towards students who achieved higher grades (A* and A) at A level, meaning that the findings cannot therefore be generalised to all students who took science A levels. There are likely to be 2 reasons for this. The first is largely a result of recruitment, with the universities who agreed to participate tending to be those with higher entry requirements for their courses. The second is methodological in that this study focuses only on those A level students who go on to study science at degree level. This approach automatically excludes those who choose not to attend university (or pick a different course), along with those who do not achieve a sufficiently good grade to be accepted on a science course at university.

The second main limitation relates to the reliability of the measurement instrument, the PSM itself. Although the PSM was designed with the intention of requiring only straightforward ‘binary’ judgements from the assessors (eg the participant has or has not done something), the process, in practice, proved to be somewhat more ambiguous in some cases. It is therefore possible that the differences in the performances of each university may reflect differences in the way assessors were standardised to apply some of the assessment criteria. In an ideal world there would have been time and resource to provide each individual assessor with a detailed ‘standardisation process’ in order to counteract the potential for slightly different standards to be applied in different universities. Although this is certainly worthy of

[11] Based on a t-test with uneven sample sizes.

consideration it is important to note that the criteria were generally not considered to be ambiguous and so this issue should not be overstated. The differences between universities may instead reflect differences in the characteristics of the students who attend them.

A similar challenge is that each participating university has a different laboratory set up (as we have discussed in the methodology section). Despite guidance to mitigate the effects of such differences, it may be that the equipment was easier to use or more familiar to a typical A level student at some universities. Indeed, there were substantial differences between the physical environments of each university which may have influenced the performance of students. A 'national test centre' where the environment and apparatus are standardised would be a method for tackling this, but this would have required a level of resource well in excess of that available and would have introduced other methodological limitations (eg arranging for a geographically diverse sample to attend the test centre).

Even if all participants were to be assessed in the same environment by the same assessors, the nature of performance assessment presents unique challenges. Laboratories are not examination halls. Sund's (2016) research illustrates some of the challenges of running performance assessments in a laboratory. For example, it can be very difficult to prevent students from watching each other if they are being assessed on adjacent work benches (although this doesn't necessary lead to improvements in performance – Sund (2016) found that students often mislead or confused each other).

5.3 Summary and interim conclusions

The evidence from this study suggests that, as intended, the reform of A level science and the changes to the assessment of practical skills have had no negative impact upon the level of practical skills that recently qualified A level students arrive at university with. In fact, those who took post-reform A levels in biology outperformed those who took the pre-reform equivalents, even when differences in achievement at A level are taken into account. There was not a substantive overall difference between the performances of the pre- and post-reform cohorts in either the chemistry PSM or the physics PSM. .

There is therefore cause for optimism that the reform is leading to a greater focus on practical work in biology, and at least not causing a reduction in the emphasis on practical work in either chemistry or physics. This is supported by the questionnaire data, where both the biology and physics post-reform cohorts reported doing practical work more frequently than the pre-reform cohort (for chemistry, the difference was less clear cut but the post-reform cohort was more likely to report doing practical work 'more than once per week' than the pre-reform cohort).

At the time of writing only a single cohort of students have completed the reformed A level qualifications in science. It will take time for teachers and students to get to

grips with the new course content and requirements, as well as the new assessment arrangements for practical work (the endorsement). It is important that Ofqual continues to monitor the new qualifications and their impact as they ‘bed in’ and become established. As part of this work, a third round of data collection will be completed for this study in September/October 2018.

6 References

Abrahams, I., & Reiss, M. J. (2015). The assessment of practical skills. *School Science Review*, *96*(June), 40–44.

Abrahams, I., Reiss, M. J., & Sharpe, R. M. (2013). The assessment of practical work in school science. *Studies in Science Education*, *49*(2), 209–251.

Abrahams, I., & Saglam, M. (2010). A Study of Teachers' Views on Practical Work in Secondary Schools in England and Wales. *International Journal of Science Education*, *32*(6), 753–768.

Acquah, D. (2013). *School Accountability in England: Past, Present and Future*. Manchester, UK: AQA Centre for Education Research and Policy.

AQA. (2016). *The new A-level science practical skills endorsement - improving science education: Applying research findings to enhance teaching and learning of STEM subjects*. Retrieved from http://filestore.aqa.org.uk/pdf/AQA-A-LEVEL-SCIENCE-STEM-RESEARCH-CONF-PAPER.PDF

Barnes, S. M., Laham, A., & Read, D. (2017). *Teachers' views of the impact of the new Chemistry A-level on students entering university in 2017/18*. Retrieved from http://edshare.soton.ac.uk/18928/13/Impact of new chemistry A-level V1.02.pdf

Biology Education Research Group. (2014). How important is the assessment of practical work? An opinion piece on the new biology A-level from BERG. *Journal of Biological Education*, *48*(4), 176–178.

Brown, C. R., & Moore, J. L. (1994). Construct Validity and Context Dependency of the Assessment of Practical Skills in an Advanced Level Biology Examination. *Research in Science & Technological Education*, *12*(1), 53–61. http://doi.org/10.1080/0263514940120107

Buchan, A. S., & Welford, A. G. (1994). Policy into Practice: the effects of practical assessment on the teaching of science. *Research in Science & Technological Education*, *12*(1), 21–29. http://doi.org/10.1080/0263514940120104

Cambridge Assessment. (2016). The Cambridge view on science. Retrieved July 19, 2016, from http://www.cambridgeassessment.org.uk/insights/the-cambridge-view-on-science/

Canning, P. (2015). How can the education system ensure students have an improved science learning experience ? *School Science Review*, *June*(96), 54–58.

Carter, I. (2014). Is this the end of the English tradition of practical A-level science? *School Science Review*, *96*(September), 12–14.

de Wolf, I. F., & Janssens, F. J. G. (2007). Effects and side effects of inspections and accountability in education: an overview of empirical studies. *Oxford Review of Education*, *33*(3), 379–396.

English, N., & Paes, S. (2015). The assessment of science practical skills : a historical perspective. *School Science Review*, *96*(June), 45–53.

Evans, S., & Wade, N. (2015). Endorsing the practical endorsement? OCR's

approach to practical assessment in science A-levels. *School Science Review*, *96*(June), 59–68.

Gatsby. (2014). New A level regulatory requirements: Response to the Ofqual consultation. Retrieved from http://www.gatsby.org.uk/uploads/education/reports/pdf/ofqual-a-level-reform-gatsby-foundation-january-2014.pdf

Gatsby. (2017). *Good Practical Science*. Retrieved from http://www.gatsby.org.uk/education/programmes/support-for-practical-science-in-schools

Gott, R., & Duggan, S. (1996). Practical work: its role in the understanding of evidence in science. *International Journal of Science Education*, *18*(7), 791–806.

Gott, R., & Duggan, S. (2002). Problems with the Assessment of Performance in Practical Science: Which way now? *Cambridge Journal of Education*, *32*(2), 183–201.

Gove, M. (2013). Letter from the Secretary of State for Education to Glenys Stacey at Ofqual. Retrieved from https://www.gov.uk/government/publications/letter-from-the-secretary-of-state-for-education-to-glenys-stacey-at-ofqual

Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, *1*(5955), 447 LP – 451.

Harlen, W. (1999). Purposes and Procedures for Assessing Science Process Skills. *Assessment in Education: Principles, Policy & Practice*, *6*(1), 129–144.

Harmon, M., Smith, T, A., Martin, M, O., Kelly, D, L., Beaton, A, E., Mullis, I, V, S., … Orpwood, G. (1997). Performance assessment in IEA's third international mathematics and science study (TIMSS). Retrieved from https://timss.bc.edu/timss1995i/TIMSSPDF/PAreport.pdf

HESA. (2017). Higher Education Statistics Agency - What do HE students study? Retrieved February 20, 2018, from https://www.hesa.ac.uk/data-and-analysis/students/what-study

Hodson, D. (1993). Against Skills - based Testing in Science Against Skills-based Testing in Science. *Curriculum Studies*, *1*(1), 127–148. http://doi.org/10.1080/0965975930010108

Hollins, M., & Reiss, M. J. (2016). A review of the school science curricula in eleven high achieving jurisdictions. *The Curriculum Journal*, *27*(1), 80–94.

JCQ. (2017). *GCE A Level Results - June 2017*. Retrieved from https://www.jcq.org.uk/examination-results/a-levels/2017/main-results-tables/a-as-and-aea-results-summer-2017

Keiler, L. S., & Woolnough, B. E. (2002). Practical work in school science: the dominance of assessment. *School Science Review*, *83*(March), 83–88.

Khan, K. Z., Ramachandran, S., Gaunt, K., & Pushkar, P. (2013). The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. *Medical Teacher*, *35*(9), e1437–e1446.

Leevers, H. (2015). If we can code a human genome, we can find a way to assess science practicals. Retrieved from https://www.theguardian.com/teacher-network/2015/feb/02/ofqual-assess-science-practicals

MacCann, R. G., & Stanley, G. (2010). Classification consistency when scores are converted to grades: examination marks versus moderated school assessments. *Assessment in Education: Principles, Policy & Practice*, *17*(3), 255–272.

Mattei, P. (2012). Market accountability in schools: policy reforms in England, Germany, France and Italy. *Oxford Review of Education*, *38*(3), 247–266.

Ofqual. (2013). Consultation on New A Level Regulatory Requirements. Retrieved from http://webarchive.nationalarchives.gov.uk/20141110161323/http://comment.ofqual.gov.uk/a-level-regulatory-requirements-october-2013/

Ofqual. (2015). GCE Subject Level Guidance for Science (Biology, Chemistry, Physics). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/447167/2015-07-20-gce-subject-level-guidance-for-science.pdf

Ofqual. (2016). *GCE subject level conditions and requirements for science (Biology, Chemistry, Physics) and certificate requirements*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/526286/gce-subject-level-conditions-and-requirements-for-science.pdf

Ofqual. (2017a). *GCE Qualification Level Conditions and Requirements*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/644330/gce-qualification-level-conditions-and-requirements.pdf

Ofqual. (2017b). Ofqual's Corporate Plan 2017-20. Retrieved from https://www.gov.uk/government/publications/ofquals-corporate-plan

Ofqual. (2017c). *The impact of qualification reform on A level science practical work - Paper 1: Teacher perspectives after one year'*. Retrieved from https://www.gov.uk/government/news/the-impact-of-qualification-reform-on-a-level-science-practical-work

Prades, A., & Espinar, S. R. (2010). Laboratory assessment in chemistry: an analysis of the adequacy of the assessment process. *Assessment & Evaluation in Higher Education*, *35*(4), 449–461.

Reiss, M., Abrahams, I., & Sharpe, R. (2012). *Improving the assessment of practical work in school science*. London. Retrieved from www.gatsby.org.uk/uploads/education/reports/pdf/ improving-the-assessment-of-practical-work-in-school- science.pdf.

Roberts, R., & Gott, R. (2006). Assessment of performance in practical science and pupil attributes. *Assessment in Education: Principles, Policy & Practice*, *13*(1), 45–67.

SCORE. (2014). *SCORE principles: the assessment of practical work*. Retrieved from http://www.score-education.org/reports-and-resources/publications-research-policy

Sloan, D. A., Donnelly, M. B., Schwartz, R. W., & Strodel, W. E. (1995). The Objective Structured Clinical Examination. The new gold standard for evaluating postgraduate clinical performance. *Annals of Surgery*, *222*(6), 735–742. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1235022/

Stacey, G. (2015). A radical experiment to end science practicals? That's just not true. Retrieved from https://www.theguardian.com/teacher-network/2015/feb/04/experiment-science-practicals-ofqual

Sund, P. (2016). Science teachers' mission impossible?: a qualitative study of obstacles in assessing students' practical abilities. *International Journal of Science Education*, *38*(14), 2220–2238.

Swanson, D. B., & van der Vleuten, C. P. M. (2013). Assessment of Clinical Skills With Standardized Patients: State of the Art Revisited. *Teaching and Learning in Medicine*, *25*(sup1), S17–S25.

The Complete University Guide. (2017). University Subject Tables 2016. Retrieved from https://www.thecompleteuniversityguide.co.uk/league-tables/rankings?o=Entry+Standards&s=Chemistry&y=2016

Toplis, R., & Allen, M. (2012). "I do and I understand?" Practical work and laboratory ise in United Kingdom schools. *Eurasia Journal of Mathematics, Science & Technology Education*, *8*(1), 3–9. Retrieved from http://ejmste.com/arsivAyrinti.aspx?kim=12

Turner, J., & Dankoski, M. (2008). Objective structured clinical exams: A critical review. *Family Medicine*, *40*(8), 574–578. Retrieved from https://www.researchgate.net/publication/23455256_Objective_Structured_Clinical_Exams_A_Critical_Review

Watts, A., & Wilson, F. (2013). *The assessment of practical science: a literature review (summarised)*. Retrieved from http://www.cambridgeassessment.org.uk/images/135793-the-assessment-of-practical-science-a-literature-review.pdf

Wellcome Trust. (2014). Wellcome Trust responds to announcement of science A-level reforms. Retrieved July 15, 2016, from https://wellcome.ac.uk/news/wellcome-trust-responds-announcement-science-level-reforms

Wellcome Trust. (2017a). *Young people's views on science education: Science Education Tracker Research Report February 2017*. Retrieved from https://wellcome.ac.uk/what-we-do/our-work/young-peoples-views-science-education

Wellcome Trust. (2017b, February 10). The Science Education Tracker. http://doi.org/10.6084/m9.figshare.4524551.v3

West, A., Mattei, P., & Roberts, J. (2011). Accountability and Sanctions in English Schools. *British Journal of Educational Studies*, *59*(1), 41–62.

Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education 3-13*, *29*(3), 17–21.

Wilson, F., Wade, N., & Evans, S. (2016). Impact of changes to practical assessment at GCSE and A-level: the start of a longitudinal study by OCR. *School Science*

Review, *98*(362), 119–128.

7 Annex A: Ofqual's A level science research programme

Reformed A level qualifications in most subjects were introduced for first teaching in September 2015 (Gove, 2013). With regard to science, the reform led to significant changes to the assessment arrangements for practical skills (Ofqual, 2016). Ofqual is conducting a programme of research to evaluate the impact of A level qualification reform on the teaching and learning of science practical skills.

The programme is comprised of 4 main studies, of which this report is Study 2:

- Paper 1: Teacher interviews – Perspectives on A level reform after one year (published)
- Paper 2: Pre and Post reform evaluation of practical ability – A comparison of science practical skills in pre and post reform cohorts of undergraduate students
- Paper 3: Valid discrimination in practical skills assessment – An exploration of classification reliability when assessing the performance of practical skills
- Paper 4: Technical functioning of assessment – An analysis of A level examination items that assess science practical skills

8 Annex B: Questionnaire for participants

To be completed by university staff:

Participant number	
University	

Practical skills research record sheet

Biology

To be completed by student:

Questionnaire

Please answer the following questions and then proceed to the first task. Please pass this record sheet to the assessor at each station so they can complete it.

1. Have you completed A-levels in any science subject (Biology, Chemistry or Physics) this year (2017)?

   Yes ☐ No ☐

2. If you have completed A-levels in Biology, Chemistry or Physics, please provide the year and grade below (please write ‘N/A’ if you did not take an A-level in the relevant subject).

Subject	**Year**	**Grade**
Biology		
Chemistry		
Physics		

3. Were any of the science A-levels you took provided by the Welsh exam board (WJEC)?

   Yes ☐ No ☐

4. Please provide details of your other *A-level* qualifications below.

Subject *eg 'Maths'*	**Year** *eg '2015'*	**Grade** *eg 'B'*

5. Please provide details of any *AS level* qualifications below.

Subject *eg 'French'*	**Year** *eg '2015'*	**Grade** *eg 'C'*

6. If you have completed any other **science** qualifications **in the last year**, please provide details below.

Qualification *eg 'BTEC'*	**Subject** *eg 'Applied Science'*	**Grade** *eg 'Merit'*

7. To what extent do you agree with the following statement: *I feel confident about carrying out practical work.*

Strongly agree	☐	Agree	☐
Neither agree nor disagree	☐	Unsure	☐
Disagree	☐	Strongly disagree	☐

8. Please estimate how often you did practical work in your school or college during your science A-levels:

More than once per week	☐	Once per week	☐
Once every 2-3 weeks	☐	Once a month	☐
Less than once a month	☐	Never	☐
Unsure / can't remember	☐		

9. Please estimate how often you did practical work in your school or college as part of your **Biology** A-level:

More than once per week	☐	Once per week	☐
Once every 2-3 weeks	☐	Once a month	☐
Less than once a month	☐	Never	☐
Unsure / can't remember	☐		

9 Annex C: Questionnaire response and PSM performance

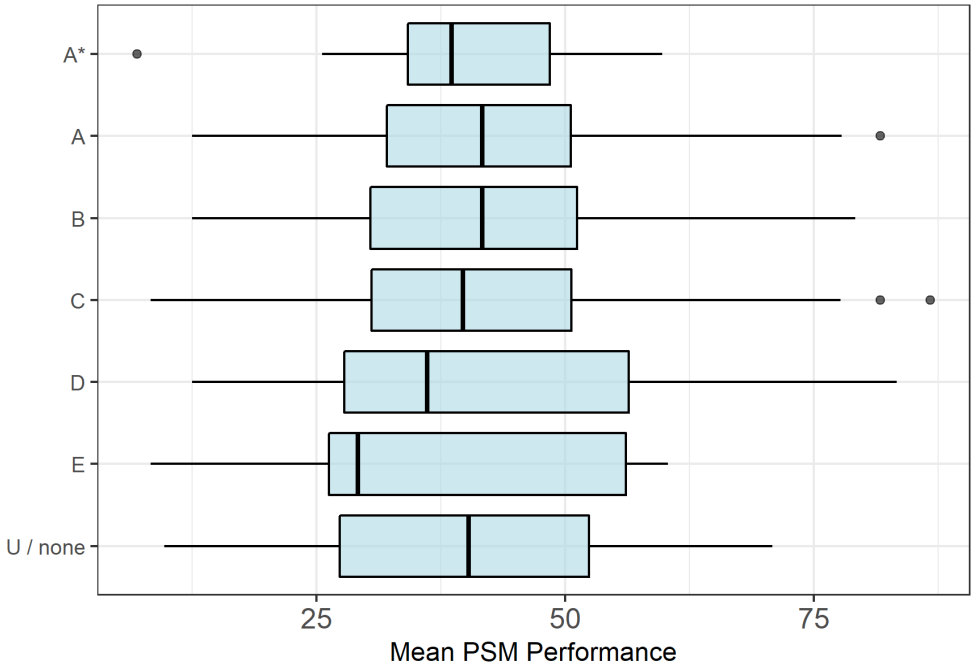

Figure 18. *Biology A level grade by PSM performance (Biology)*

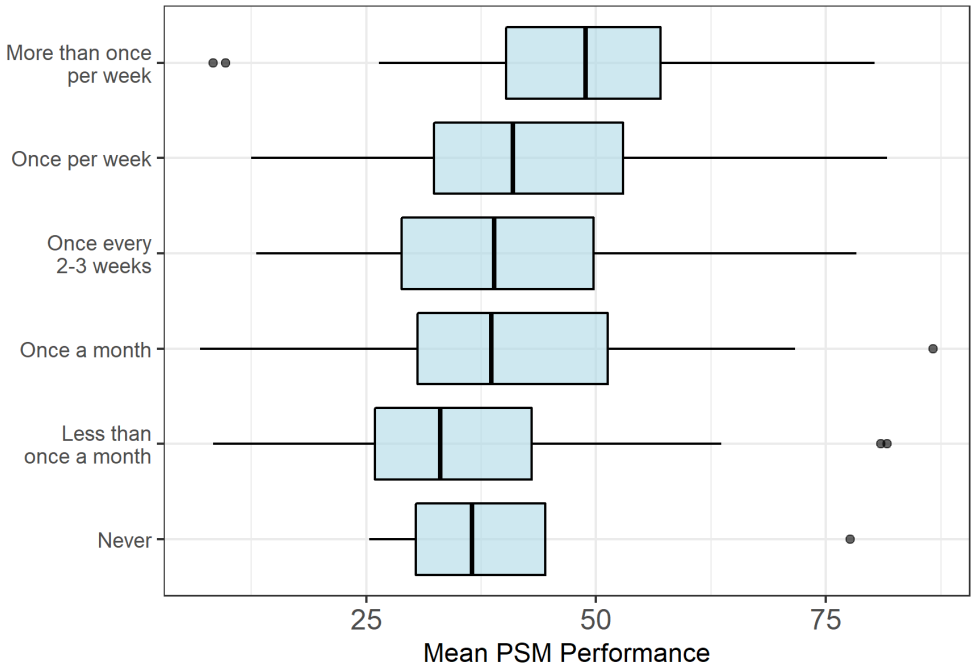

Figure 19. Participants' self-reported frequency of practical work at A level by PSM performance (Biology).

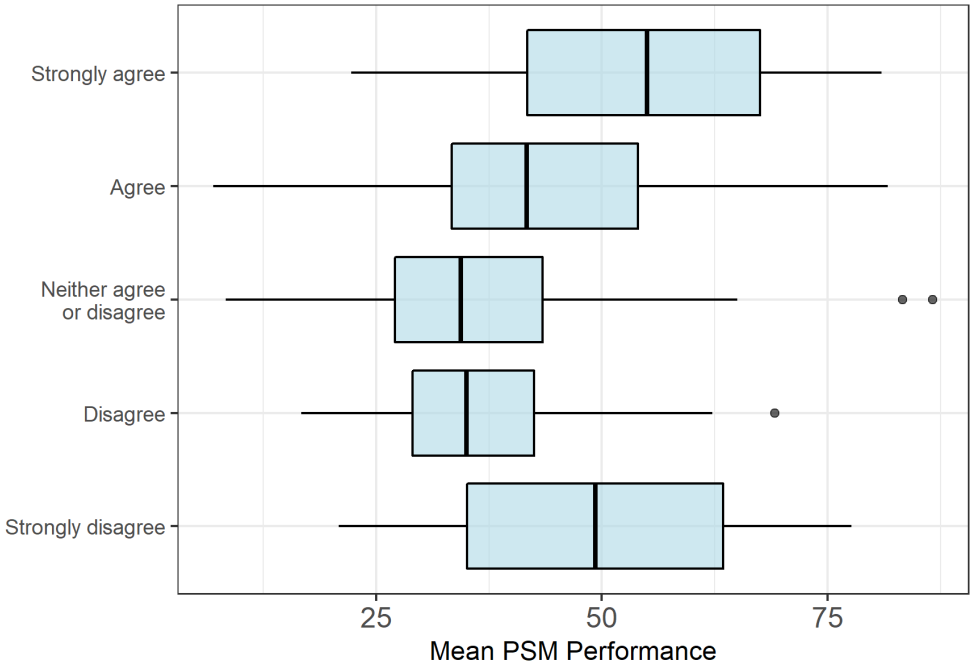

Figure 20. *Participant confidence by PSM performance (Biology).*

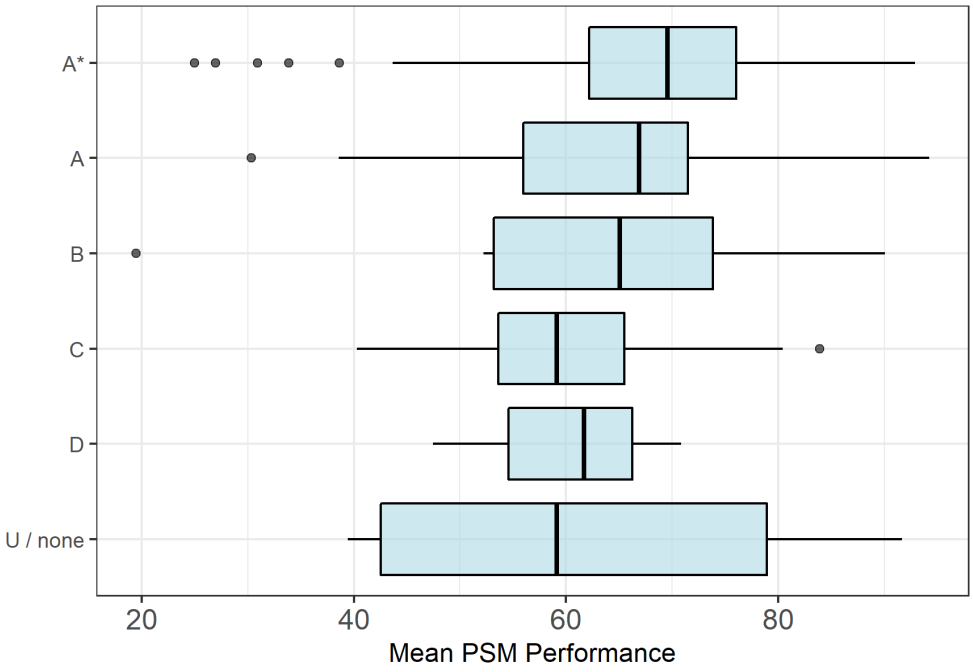

Figure 21. Chemistry A level grade by PSM performance (Chemistry)

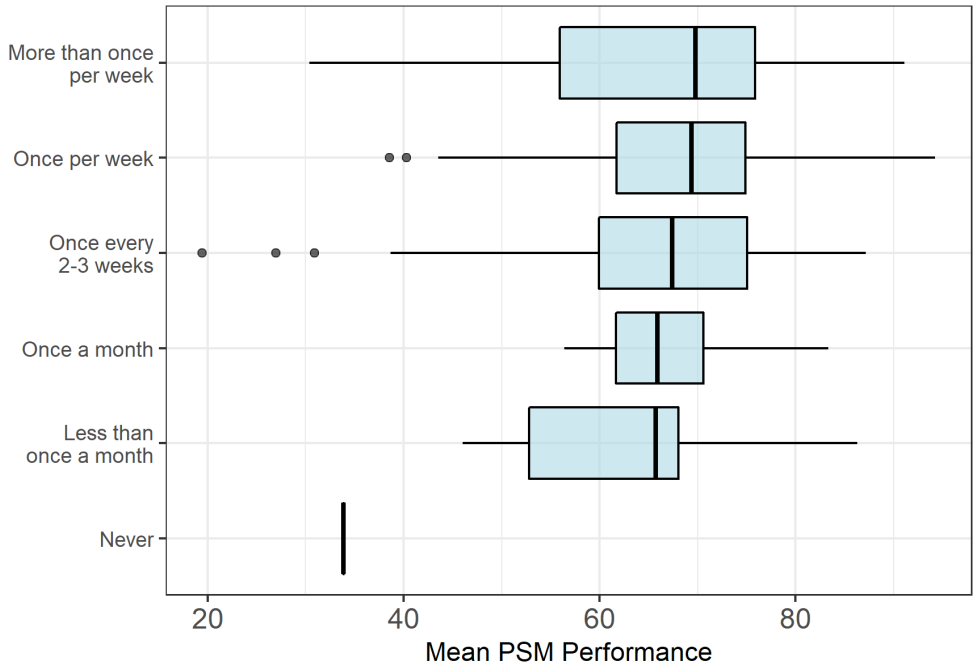

Figure 22. *Participants' self-reported frequency of practical work at A level by PSM performance (Chemistry).*

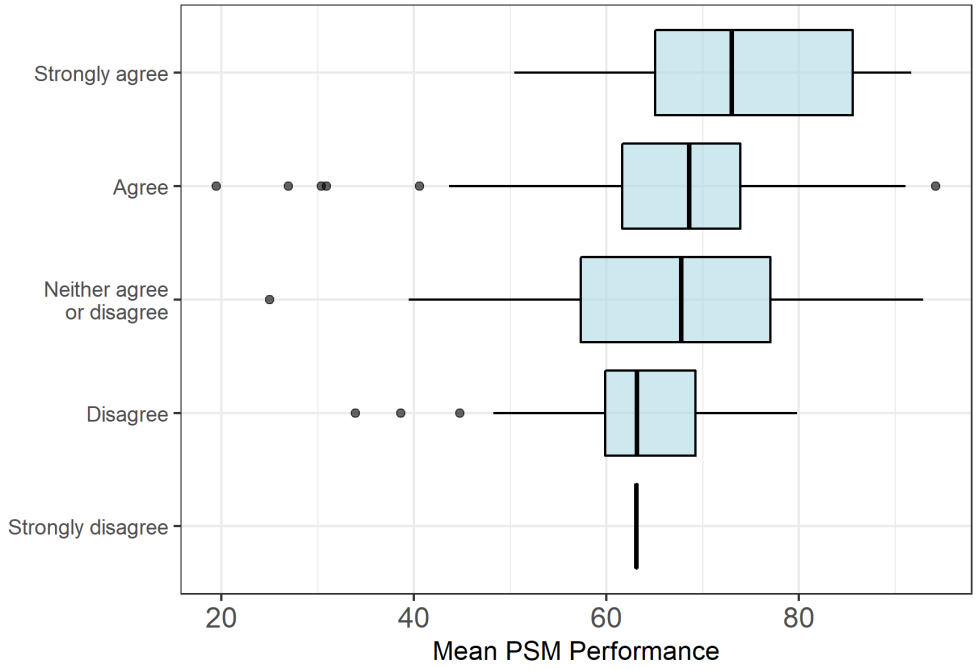

Figure 23. *Participant confidence by PSM performance (Chemistry).*

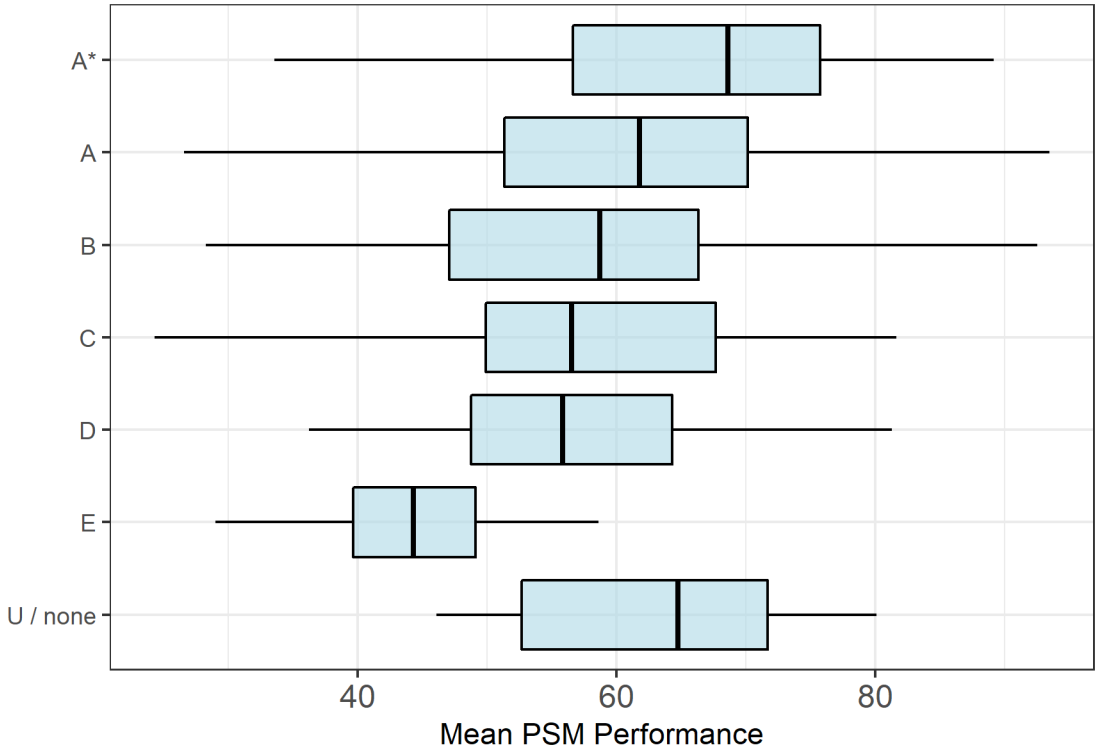

Figure 24. *Physics A level grade by PSM performance (Physics)*

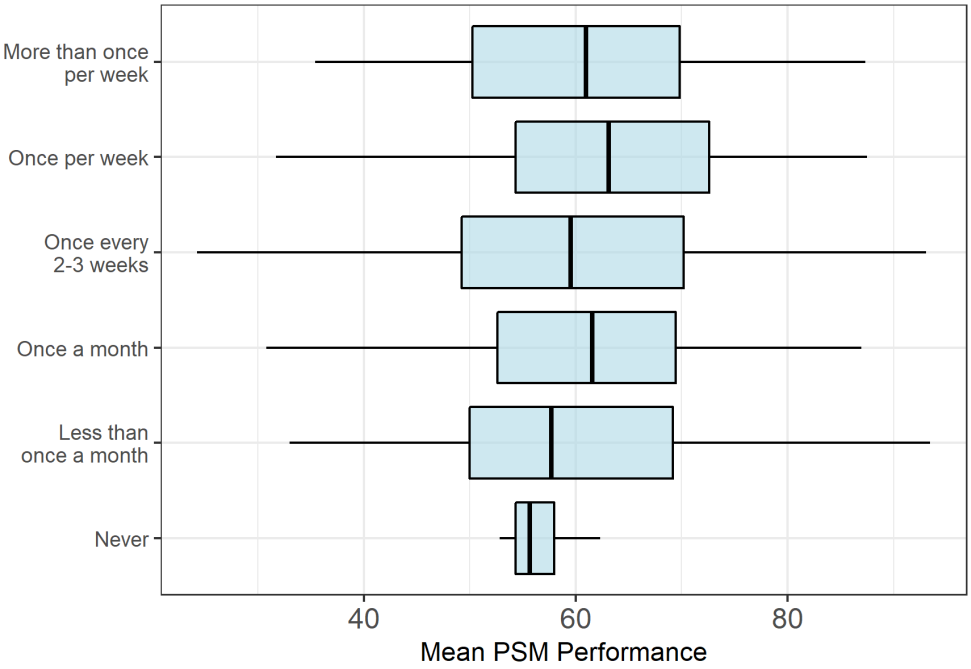

Figure 25. *Participants' self-reported frequency of practical work at A level by PSM performance (Physics).*

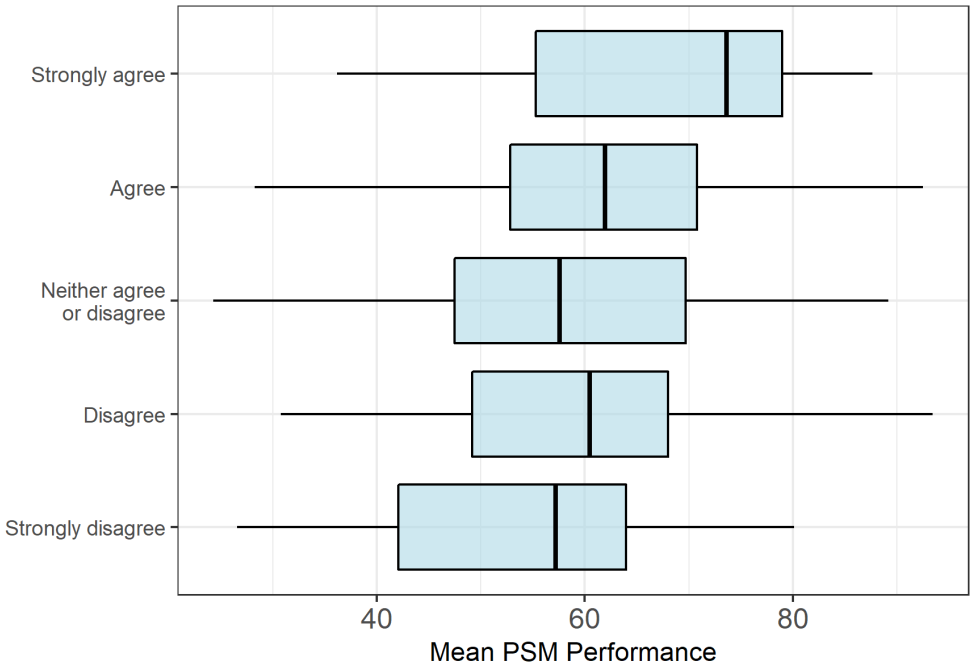

Figure 26. *Participant confidence by PSM performance (Physics).*

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344