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

Paper 3: Valid discrimination in the assessment of practical performance

May 2018

Ofqual/18/6372

Authors

This report was written by Stuart Cadwallader, Benjamin M P Cuff and Aneesa Khan, from Ofqual's Strategy, Risk and Research directorate.

Acknowledgements

The authors gratefully acknowledge the support and expertise of Dr Tasnim Munshi and the Chemistry department at the University of Lincoln. The authors are also very grateful for feedback received from members of Ofqual's Research Advisory and Standards Advisory groups.

Contents

1 Executive summary

A level qualifications in science have recently undergone reform and there have been significant changes to how practical skills are assessed. One such change requires teachers to complete direct assessment of practical work, assessing students on a minimum of 12 practical activities which include the use of specific apparatus and techniques[1]. Students are expected to build towards competence in practical work throughout the duration of the course in order to pass what is known as the 'practical endorsement'. The endorsement does not contribute to the primary A level grade, instead each student receives a separate result; either a 'Pass' (for meeting each of 5 assessment criteria) or a 'Not classified'. In addition to this direct assessment, students' practical knowledge and skills[2] are also assessed indirectly through examination questions (a minimum of 15% of the available examination marks must be allocated to the assessment of practical skills).

The reform has highlighted a number of considerations with regard for the assessment of practical work. One such question is whether the direct assessment of performance in practical work is best conceptualised in binary terms (eg the student is either competent or they are not) or whether a more nuanced grading structure (whereby multiple grades are available) could be more valid. To begin to investigate this question, Ofqual has conducted a study to explore the extent to which subject experts can consistently discriminate between practical performances using different holistic grading scales. Judgements about performance were of an overarching 'best fit' nature, rather than based on prescriptive assessment criteria (eg there was not a detailed marking scheme). Through this approach, the intention is to ascertain whether there is a particular number of grades into which examiners can best 'naturally' allocate practical performances in a reliable manner.

It is important to pause here and note that this study does not evaluate the current assessment arrangements, which take a far broader approach to assessing practical skills (eg the assessment criteria relate to competencies that must be evidenced over the duration of the course, they do not focus on competence in specific techniques as performed on specific tasks). There are broader questions about what the direct assessment of practical work in national qualifications should achieve, discussions that lie beyond this scope of this paper. However, it is hoped that the study could inform such discussions should further reform be instigated in the future.

---

[1] Please see Appendix 5b of the of the Department for Education's (2014) subject content for details of the skills and knowledge that are to be covered in the examinations.
[2] Please see Appendix 5a of the of the Department for Education's (2014) subject content for details of the skills and knowledge that are to be covered by the Practical Endorsement.

To begin to investigate this question, Ofqual has conducted a study to explore the extent to which subject experts can consistently discriminate between practical performances using different holistic grading scales. Judgements about performance were of an overarching 'best fit' nature, rather than based on prescriptive assessment criteria (eg there was not a detailed marking scheme). Through this approach, the intention is to ascertain whether there is a particular number of grades into which examiners can best 'naturally' allocate practical performances in a reliable manner

Fourteen examiners were recruited to assess the performance of students as they undertook short practical tasks in chemistry. Performance of this kind is ephemeral, so examiners assessed pre-recorded video footage, thus ensuring that they were all assessing the exact same set of performances. A repeated measures design was employed to compare the consistency of examiners' judgements across 4 experimental conditions, each of which required the application of a different approach to grading. These conditions were: 'Fail/Pass', 'Fail/Pass/Merit', 'Score of 1 to 5' and 'Score of 1 to 10'. The dependent variable was the level of agreement within each experimental condition, which was statistically controlled to account for 'chance' agreements in classification.

The findings suggest that, for the practical activities that were selected, unguided holistic judgements are not particularly reliable across subject experts, even when only 2 grades are available. Elements of the task and the performance interact to create 'grey areas', ambiguities that can lead to inconsistency between judges. A carefully constructed marking scheme or list of assessment criteria, probably combined with a robust system for standardisation, would likely be necessary should individual practical techniques be assessed in isolation. Such criteria would need to be bespoke for the practical task in question and detailed enough to account for the different types and qualities of performances that may be observed. This is not necessarily a surprising finding, but one worth reflecting upon.

Despite this, the study did find some limited evidence with regard to how many grades may be appropriate for assessing tasks of the type that were included in the study. For example, even when the effects of chance agreement are taken into account, examiners were just as reliable in discriminating between 3 grade levels ('Fail/Pass/Merit') as between 2 grade levels ('fail/pass'). However, examiners were less able to reliably apply the 5- and 10-point grade scales, suggesting that these scales may be too granular as the basis for holistic judgements about performance. The findings may also tell us something about the nature of examiners underlying judgements when they assess practical performance, suggesting that their concept of competence (and their application of this concept) may be important. This report explores and discusses the findings, including limitations and potential applications, in further depth.

2 Introduction

Reformed A level qualifications in biology, chemistry and physics were introduced for first teaching in September 2015. As part of this reform, the assessment arrangements for practical work have changed significantly (see Annex A for details). Ofqual is conducting a programme of research to evaluate the impact of these changes on students' practical skills (see Annex B). This report reflects one part of this programme, focusing on valid and reliable discrimination when assessing the *performance* of practical work. The study reported here explores the degree of granularity to which the performance of practical skills can be reliably assessed. It considers whether subject experts can consistently differentiate across a spectrum of performance quality or whether it is more appropriate to apply binary judgements (eg Pass/Fail). This report aims to enhance our theoretical understanding of the holistic assessment of performance and therefore serve to inform our future thinking about relevant assessment arrangements.

2.1 Direct assessment of practical work in the reformed A levels

Though this report will take a fairly broad and theoretical approach, it is important to set it within the context of recent qualification reform. For post-reform A level science qualifications, teachers are required to complete assessment of practical work throughout the duration of the course, assessing the performance of students across a minimum of 12 practical activities which, taken together, cover the use of specified techniques and apparatus[3]. Students are expected to have developed competency against 5 Common Practical Assessment Criteria (CPAC), so named because they are common across biology, chemistry and physics (see Ofqual, 2016). It is not necessary for each practical activity to assess all 5 of the CPAC, but students must be given sufficient opportunity to develop and demonstrate competency in each by the end of their course. The CPAC are as follows:

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

[3] The prescribed techniques and apparatus can be found in Appendix 5c of the subject content (Department for Education, 2014).

5. Researches, references and reports

Schools and colleges are required to curate evidence of success against the CPAC for each of their students. They must demonstrate to exam boards that they are conducting the required practical work and that the assessment is being undertaken correctly. There are 2 possible outcomes available for this 'practical endorsement' (which is reported as a separate result to the student's A level grade); the student either receives either a 'Pass' or a 'Not Classified'. It is important to note that students' knowledge and understanding of practical skills[4] are also assessed via the written examinations that take place at the end of the A level course. A minimum of 15% of the available examination marks must be allocated to the assessment of practical skills (Ofqual, 2015).

These assessment arrangements are intended to have a positive impact on teaching and learning, encouraging a formative process whereby students work towards achievement of the CPAC. Qualitative research about teachers' initial experience of the reformed qualifications suggest that the new assessment arrangements may encourage a greater focus on the development of skills and allow teachers to better integrate practical work with the course content (Ofqual, 2017). However, the findings also suggest that the mechanisms for achieving this are dependent on the precise context of the school or college, meaning that the impact of qualification reform is unlikely to be uniform across all institutions. The medium and long term 'wash-back' effects that the assessment arrangements will have on teaching and learning are yet to be revealed and, as recommended in a recent report from the Gatsby Foundation (2017), it will be important to monitor these carefully.

The question of how many outcomes (grades) should be available for the endorsement first came to the fore following Ofqual's initial consultation on the new assessment arrangements (Ofqual, 2013). Correspondence from Dr Anne Scott and Dr Elizabeth Swinbank of the University of York Science Education Group (YSEG) suggested that many key practical skills could be graded as either 'Fail', 'Pass' or 'Distinction' (with distinction generally awarded to students who demonstrate the ability to make a decision about their actions based on an underlying scientific principle). YSEG suggested that some skills would probably not be appropriate for the application of 3 grades and could only be reliably assessed on a Pass/Fail basis (for example, skills such as 'following multi-step instructions').

Ofqual considered this element of the recommendation but decided to apply a binary grading approach to all 5 CPAC (Ofqual, 2016). The main reason for this decision

[4] Please see Appendix 5b of the of the Department for Education's (2014) subject content for details of the skills and knowledge that are to be covered in the examinations.

was that a 'Pass/Fail' approach was considered to be much more manageable and straightforward for schools and colleges to use reliably. The educational benefits of having additional grades in the assessment were not sufficiently clear to warrant the increased complexity.

This report considers grading further by exploring the nature of assessment decisions about performance. Specifically, it explores binary competence judgements in the context of assessing the performance of individual practical techniques and skills, which are in essence the 'milestone' decisions that lead to a student's final outcome in the practical endorsement. In particular, these performance-related judgements are required in relation to the first 3 CPAC, the assessment of which require the teacher to directly observe the student performing 'hands-on' practical work. It may be that the binary approach to grading is not fine grained enough, failing to reflect reliable holistic decisions that teachers or examiners would naturally make when assessing the competency of a performance.

It is important to emphasise that this is not a direct evaluation of the suitability of the CPAC or the approach to assessing the practical endorsement. For the endorsement, judgements about competency regard a range of activities over time. The CPAC relate to broad competencies that must be evidenced by the end of the course, they do not focus on competence in specific techniques and against specified activities. It is important to discuss some of the nuances of defining and assessing practical work, and clarifying the focus for this study, before elaborating further.

2.2 Performing practical work

Practical work draws upon a considerable breadth of knowledge and a wide range of skills[5]. The Science Community Representing Education (SCORE, 2014) suggest that educationally effective practical work should include the following 3 elements: the development of conceptual understanding, opportunities for extended investigation (including planning, analysis and evaluation), and training in technical and manipulative skills. This study focuses mainly on the third of these elements.

To clarify, Abrahams, Reiss & Sharpe (2013) define practical skills as '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' (p. 210). This is an operational definition, which deliberately focusses on technical and manipulative skills, excluding the conceptual understanding that may be required for certain elements of practical work (eg

[5] There is some ambiguity regarding precisely what is encompassed by the term 'practical work' and so there is no universal definition (Lunetta, Hofstein, & Clough, 2007).

planning an investigative experiment or writing a laboratory report). The emphasis is on the *physical performance* of practical work.

Such technical and manipulative skills appear to be of considerable importance to both teachers and employers. Science, Technology, Engineering and Maths (STEM) employers often define the practical skills they value in terms of 'dexterity, 'hand-skills' and 'lab work' (Gatsby, 2012, p. 3). The ability to 'carry out standard procedures and tests' and 'take and record observations and measurements with accuracy and precision' are particularly valued and sought in school leavers (Gatsby, 2012, p. 8). A recent survey of A level science teachers suggests that, while there are many reasons for undertaking practical work in the classroom, the top 3 were to: 'develop manipulative skills and techniques', 'develop reporting, presenting, data analysis and discussion skills' and 'encourage accurate observation and description' (Wilson, Wade, & Evans, 2016)[6].

2.3 Assessing performance

Arguably, the most valid approach to assessing a student's practical skills is to observe them engaging in a relevant practical task. Abrahams, Reiss, & Sharpe (2013, p. 245) suggest that 'whilst a conceptual understanding of the topology of knots and manifolds might well be assessed by a written task, the most effective means of assessing whether a student is competent in tying his/her shoe laces is, we would argue, to watch him/her as he/she attempt to tie them'. They suggest that such skills are best assessed *directly*, via observation.

The conceptual distinction between Direct Assessment of Practical Skills (DAPS) and Indirect Assessment of Practical Skills (IAPS) is useful here (Abrahams et al., 2013). IAPS refers to assessment which seeks to infer the student's level of proficiency through the scrutiny of an artefact that they have produced (eg a written report, table of data, or graph), while DAPS refers to assessment which requires the student to demonstrate their proficiency through the physical manipulation of apparatus and the execution of specific techniques.

Though DAPS is an intuitively more valid approach to assessing performance there are unique challenges if the goal is reliable and replicable assessment. This is because the actual physical performance of a technique or process is ephemeral – it cannot be perfectly captured for re-assessment or moderation at a later date, unlike the outputs or artefacts that may have been produced by that performance.

[6] It is worth noting that similar research with teachers of key stage 3 students (11-14 year-olds) suggested that the main reason for conducting practical work was to generate and maintain interest in science (Abrahams & Saglam, 2010). It may be that A level students, who have chosen to continue their study science, are not perceived to require this type of encouragement to the same extent.

For both DAPS and IAPS, it can be difficult to ascertain exactly what is being assessed in the performance of a particular practical task. Gott & Duggan (2002) suggest that technical and manipulative skills are likely to rely on some degree of 'procedural understanding' with regard to how to collect and validate scientific evidence, suggesting that such skills may therefore be 'inseparable in practice from the conceptual understanding that is involved in learning and applying science' (Harlen, 1999, p. 129). It seems reasonable to suggest that most practical tasks will rely on some theoretical understanding of the phenomena which is to be investigated or else the activity would be meaningless. Technical and manipulative skills are therefore difficult to assess in isolation because related (but separate) skills and knowledge are likely to be implicit in the performance.

2.4 Competency, proficiency and valid discrimination

At this point, it is worth explaining how the terms *competency* and *proficiency* will be used in this report. With regard to the assessment of performance, the term competency is often deployed in a binary manner, with an individual judged on whether or not they possess a desirable skill. For example, a candidate can either successfully set up a burette (they pass) or they cannot (they fail). To return to our earlier example, a person can either tie their shoe laces or they cannot.

Such terminology is straightforward to interpret and could provide a valid way of categorising performances on particular tasks. However, a binary approach may also be an over-simplification that masks a more valid conception; that a candidate's performance may be placed somewhere along a spectrum of proficiency for the task. In this context, the term competency may be used to refer to something slightly different, specifically a benchmark located somewhere along a proficiency spectrum which represents the minimum acceptable level of proficiency that the candidate should exhibit. The assessment outcome for the candidate may still be binary (pass or fail) but it may be underpinned by a more nuanced judgement as to where the candidate is located on the spectrum (and perhaps where the benchmark for competency should be placed). For example, perhaps a student can successfully and safely operate a piece of scientific apparatus but could be more proficient in their technique. Perhaps they could be more efficient or achieve more accurate readings.

Competency can also be conceptualised multi-dimensionally. For example, vocational competency is often assessed against a continuous spectrum which comprises various components of knowledge, skill and behaviour – it is a holistic judgement based on multiple factors. In the construction industry, an individual is considered to develop their competence (which may increase or decrease) over the course of their career (Pye Tait, 2014). Such nuance may also apply in the context of science practical skills, particularly where a task or performance could be considered to be multi-dimensional (for example, where a practical technique requires multiple

steps or techniques). This study sets out to explore some of these nuances of judgement and how they are used to differentiate between performances.

The reformed A level science qualifications require teachers to directly assess competency across a range of practical techniques. In this context, a candidate may be described as demonstrating competence if they are able to surpass an established benchmark of proficiency (eg a titration is conducted safely and accurately) by the end of their course. This raises an interesting question about the extent to which varying degrees of proficiency can be reliably discriminated with regard to the performance of practical work in science. Subject experts may generally agree in their judgement about whether a particular candidate has demonstrated sufficient proficiency to be deemed competent (eg Fail/Pass). However, proficiency may also be reliably discernible across a wider spectrum (for example, using a 5 point scale), allowing a greater degree of differentiation between candidates.

With too few scale points, the assessment would have limited ability to discriminate between candidates of different abilities, therefore impacting on the utility of assessment outcomes for selection purposes (eg for university admissions officers or employers wishing to select the most appropriate candidates). With too many scale points, markers may not be able to meaningfully and reliably distinguish between neighbouring scale points, impacting on the validity of assessment outcomes (MacCann & Stanley, 2010; Newton, 2007).

2.5 Research objective

This study seeks to investigate the extent to which direct assessment of practical performance in chemistry can reliably discriminate across scales of differing length. Chemistry has been selected because the subject content emphasises a number of practical techniques (Department for Education, 2014, p. 22), many of which are relatively easy to prepare for direct assessment. It should be noted, however, that meaningful differentiation when assessing performance is a question for a number of subjects and qualifications, not just chemistry and the sciences. The study sets out to answer the following research question:

> *When assessing the performance of candidates' undertaking specific practical activities, to what extent does inter-rater reliability vary when differentiating across rating scales of 2, 3, 5 and 10 scale points?*

The study is explorative in that it seeks to enhance our understanding of how subject experts make holistic judgements about practical performance and whether consistent judgements about proficiency can be made across scales of varying gradation. The findings will provide evidence to help inform assessment developers about the optimum number of scale points into which it is meaningful and reliable to classify practical performance.

To reiterate an earlier point, this study does not seek to directly evaluate current assessment arrangements regarding practical skills at A level. Assessing a specific practical performance is different to assessing success against the CPAC over the duration of a 2-year course. The CPAC are intentionally broad and overarching in nature, while assessing a single performance requires a considerably more focussed and summative judgement. Even so, assessing performances on particular tasks underpin the CPAC decisions, so it is important to understand how such judgements are approached and the degree to which they may discriminate most reliably.

3 Method

3.1 Design overview

Fourteen experienced chemistry examiners were recruited to assess 5 separate 'mock candidates' on each of 4 different practical tasks. A repeated measures design was employed to compare the consistency of examiners' judgements across 4 experimental conditions, each of which required the application of a different rating scale for the assessment of performance. The dependent variable was the mean level of agreement within each experimental condition. Supporting information was gathered to further explore the nature of the examiners' judgements.

While written examinations produce evidence that can be validly assessed by multiple examiners (exam scripts), the performance of practical skills is ephemeral. In order to ensure that all participants were assessing the exact same performances, they viewed pre-recorded video footage of the tasks being undertaken. The 4 tasks and 4 rating scales (ie the experimental conditions) are described in further detail below, and in Annex C.

3.2 Materials

This study made use of 4 chemistry practical tasks[7], all of which require the use of techniques and apparatus that are compulsory to the A level chemistry course. Detailed information about these tasks is provided in Annex C, but they can be summarised as follows:

- Task 1: Setting up a burette
- Task 2: Thin layer chromatography
- Task 3: Setting up a reflux and distillation

[7] The 4 tasks were designed by subject experts as part of a separate research study (please see Annex B for an overview of the research programme).

- Task 4: Making up a standard solution

A 'mock' candidate (a postgraduate chemistry student from the University of Lincoln) was filmed undertaking each task 5 times, changing their approach in a deliberate way for each of these performances. In broad terms, the performances could be classified as: a good performance (no errors, n = 4), a mixed performance (some minor[8] errors, n = 5), a poor performance (minor and significant[9] errors, n = 4), a performance with one significant error (n = 5), and a creative approach (a good performance which uses an unusual but technically acceptable approach, n = 2). The repetitions were designed by a senior academic from the University of Lincoln's chemistry department and were based on common errors or misunderstandings that had been observed in undergraduate chemists.

The researcher who operated the camera followed a plan devised by the senior academic which ensured that, as far as possible, each performance was captured from the same angles and under the same lighting conditions. This generally required a static camera angle but occasionally required panning or zooming to focus the frame on a particular aspect of the performance (eg the measurement scale on the side of a burette).

The 4 experimental conditions varied by the rating scale that was employed. The 4 scales and the accompanying guidance are described in Table 1.

Table 1. *The 4 rating scales (experimental conditions).*

Rating scale	Available grades	Guidance for participants
2 grades	Fail/Pass	A 'Pass' should indicate that, in your professional judgement, the candidate has exhibited a sufficient level of performance to be considered competent at the task.
3 grades	Fail/Pass/ Merit	A 'Pass' should indicate that, in your professional judgement, the candidate has exhibited a sufficient level of performance to be considered competent at the task. A 'Merit' should indicate that the candidate has exhibited a high quality of performance, clearly surpassing what you would consider to be the 'Pass' standard.

[8] Minor errors were those which may reduce the accuracy of results or represent suboptimal practice.
[9] A significant error was defined as one which caused a health and safety issue or would lead to the task producing invalid results.

5 point scale	1 to 5	How would you rate the student's performance on a scale of 1-5 (with 1 indicating a poor performance, and 5 indicating an excellent performance)?
Ten point scale	1 to 10	How would you rate the student's performance on a scale of 1-10 (with 1 indicating a poor performance, and 10 indicating an excellent performance)?

In order to neutralise the possible practice and/or fatigue effects associated with repeating the assessment for each of the 4 conditions, the order in which participants viewed each task and each of the 5 videos (the performances) within that task were randomised. Participants applied all 4 rating scales (in a randomised order) to each performance before moving on to the next one.

3.3 Participants

A level chemistry examiners were invited to participate in the study on the basis of a recommendation from their parent exam board[10]. Exam boards were asked to recommend a pool of examiners who met the following criteria:

- a current teacher of A level chemistry
- a minimum 3 years of *teaching* experience in A level chemistry
- a minimum 2 years of *examining* experience in A level Chemistry
- experience of moderating the national assessment of practical skills (desirable but not essential)

A level chemistry examiners were recommended and 14 agreed to participate in the study. Participating examiners were paid a fee for their involvement and all were provided with the same window of time in which to complete the work remotely.

3.4 Procedure

Examiners completed the study using a combination of an online survey website and email. They completed questionnaires about each video[11], rating them using each of the 4 rating scales and noting anything which they felt was relevant to their decision. The order in which each of the twenty videos were assessed and the rating scale

[10] The AQA, OCR, Pearson and WJEC exam boards participated in the study.
[11] The videos were hosted on a private channel of a video-sharing website and accessible only via a hyperlink that was embedded in each online questionnaire.

which was used was randomised, as dictated by an examiner-specific 'Completion and Ranking' grid[12].

By way of support, examiners were provided with outlines and 'performance indicators' for each of the 4 practical tasks (along with the aforementioned guidance about each grading scale). The performance indicators identified general features of the performance which they may wish to look for when exercising their judgement. Examiners were not provided with a mark scheme to help map the performance to rating decisions. Instead, they were asked to exercise their own judgement. The outlines and performance indicators for the 4 tasks are provided in full in Annex C.

Once the rating process had been completed for all 5 videos from a particular task, examiners were asked to place the performances in rank order based on their judgements of their relative quality. This was repeated for each of the 4 tasks. Finally, examiners completed an 'exit questionnaire' which captured their opinions about each rating scale. Examiners were invited to provide reasons for their decisions and to specify which of the 4 rating scales they preferred.

Before continuing, the deliberate absence of a strong framework for guiding judgements (eg a mark scheme) warrants further explanation. It would have been possible to develop mark schemes for each of the 4 rating scales. For example, it would have been a relatively simple task for a subject expert to develop a 10 point mark scheme to credit a student for achieving each of 10 prescribed actions on a given task (eg creating criteria such as 'replaces the stopper on the test tube' and 'adds the correct volume of the solution'). Assuming that these points were reasonably clear and unambiguous, the reliability of this mark scheme (the consistency of scores across different judges) would probably be quite high.

However, this type of 'penny point' marking may not reflect the genuine quality of the performance in a holistic sense – the outcomes may lack validity. As an alternative, a 'levels of response' mark scheme may be developed in which judges are required to match the performance they observe to one of several level descriptors using a 'best-fit' approach. This would involve a more holistic judgement but would still require design decisions regarding the number and width of the levels – decisions which can have an impact on marking consistency (Pinot de Moira, 2011).

Writing a mark scheme requires the author to make substantive judgements about what the assessor should value in a student's performance (Ahmed & Pollitt, 2011) and decisions about how best to assess what they consider to be important. Instead,

[12] There were 3 levels to the randomisation: the order in which examiners viewed the task videos, the order in which examiners viewed the performances for each task, and the order in which examiners applied each of the 4 rating scales to each video.

this study seeks 'proof of concept' that scales of 3 points or more provide a framework around which examiner judgements can reliably form. By giving minimal guidance to examiners, this study attempts to elicit more fundamental judgements about the quality of performances in order to establish whether they naturally fit with any of the grading schemes. Such an approach would be inappropriate for a high stakes assessment (such as A level), for which a high degree of consistency across assessors would be required.

4 Results

4.1 Inter-rater agreement

Table 2 shows the degree to which the examiners agreed in their judgements using each rating scale (the degree of inter-rater reliability). Percentage agreement statistics show the propensity for perfect agreement between pairs of examiners. Findings show that as the scale length increases, agreement appears to decrease, though it is apparent that none of the rating scales elicited high levels of agreement.

Percentage agreement statistics do not take into account the fact that scales of different length are differentially effected by chance ratings (guessing). Two judges who are grading randomly are more likely to agree on a grade by chance when there are fewer grades from which to choose[13]. 'Kappa statistics' allow one to partially circumvent this issue, as they attempt to account for chance agreement. One such coefficient, which has been shown to be relatively robust, is the Gwet coefficient (Gwet, 2008)[14]. This can be defined as the conditional probability that 2 randomly selected examiners agree if there is no agreement by chance (Blood & Spratt, 2007).

[13] Theoretically, test items with fewer scale points are always marked more accurately (Pinot de Moira, 2013). With regards to grading, classification consistency is also higher when fewer grades are available (MacCann & Stanley, 2010).

[14] The 'AC1' version of the coefficient is used for this analysis. This 'first order agreement' statistic credits only 'perfect agreement' between examiners. It is possible to apply weightings that adjust for varying magnitudes of disagreement (Blood & Spratt, 2007). The AC2 is a 'second order agreement' statistic that adjusts for such weightings. However, this approach requires the researcher to make a judgement about the extent to which low magnitude disagreement should be 'partially credited'. Given the lack of a theoretical basis on which to make such a judgement, weightings were not applied.

Table 2. *Inter-rater agreement statistics for the 4 rating scales and rank ordering.*

Grading scheme	% Agreement	Gwet AC1	95% LCL	95% UCL
2 grades	72.86	0.48	0.27	0.70
3 grades	62.03	0.48	0.29	0.68
5 point scale	39.84	0.25	0.13	0.38
Ten point scale	19.95	0.11	0.03	0.19
Rank within test	*35.76*			

When chance agreement is taken into account, examiners' ratings appear to be equally reliable on the 2- and 3-grade scales. However, the 5- and 10-point scales still appear to be less reliable than the 2 and 3 point rating scales. The coefficients fall below the 0.4 threshold, which is often suggested as a minimum benchmark for moderate agreement (eg Wongpakaran, Wongpakaran, Wedding, & Gwet, 2013). This finding is weakened by the confidence intervals, which are very wide for the 2 and 3 grade rating scales in particular and have lower limits that are below the 0.4 threshold. For the 2 and 3 grade scales, the uncertainty reflects the relatively low levels of agreement, which, as we shall see, are focussed around the assessment of particular performances.

Finally, it is worth noting that percentage agreement on the rank of each performance (within each task) was only around 36%. However, the Spearman correlations between examiners rankings were largely moderate or strong (0.4 to 1.0) and statistically significant ($p < .01$).

4.2 Examiner preferences

Examiners were asked which rating scale they thought was best for assessing the students' performance and why. Of the 14 examiners, 6 preferred the 3 grade scale (fail/pass/merit), 6 the 10 point scale and 2 the 2 grade scale (Fail/Pass). None of the examiners favoured the 5 point scale. Typically, those who preferred the 10 point scale felt it allowed them to better differentiate between performances, something which they felt would be of educational benefit to learners.

> It would allow a student to see where they stand on a scale and where there is room for improvement. The Pass/Fail is a little restrictive.
>
> *Examiner 10 – regarding 10 point scale*
>
> Scoring system is good as it is generally possible for learners to achieve some credit.
>
> *Examiner 5 - regarding 10 point scale*

Those who preferred the 2 and 3 grade scales generally cited less ambiguity in decision-making and an ability to make a more holistic judgement as the reason for their preference.

> It is clear cut - have they achieved the standard or have they not? The merit system allows best practice to be recognised once the basics have been covered.
>
> *Examiner 14 - regarding Fail/Pass/Merit point scale*

> The criteria were explained and give a holistic assessment. It is difficult to assign a number- what is the difference between 5 and 6?
>
> *Examiner 9 - regarding Fail/Pass/Merit point scale*

There was a tendency for those who preferred the 3-grade scale to explain their preference in terms of assessment quality: the precision and ease of their judgements. Those who preferred the 10-point scale seemed to have feedback to the learner in mind, whether that be to help them to improve or to provide them with credit for particular elements of their performance.

4.3 Performance type and task

The previous section presents the overall inter-rater reliability across all 20 performances. In actual fact, reliability varied considerably between performances. It is possible to break the data down by the 'type' of performance demonstrated in each video that was assessed (see page 9). Table 3 displays the percentage agreement and Gwet coefficients for each of the 4 rating scales across each of the 5 performance types. The confidence intervals for the Gwet coefficients are too wide for strong conclusions to be drawn but, in general, the 'Poor' performances seem to elicit the highest inter-rater agreement, followed by performances which feature a single significant error.

Table 3. *Reliability coefficients, grouped by performance type*

		2 Grades	3 Grades	5-point scale	10-point scale
Good (n = 4)	% Agreement	76.37	40.66	29.95	15.11
	Gwet AC1	0.69	0.15	0.15	0.05
	95% LCL	0.14	-0.05	0.03	0.01
	95% UCL	1	0.36	0.27	0.09
Creative (n = 2)	% Agreement	68.13	65.93	29.67	15.38
	Gwet AC1	0.41	0.55	0.13	0.05
	95% LCL	-5.98	-3.52	-1.72	-0.83
	95% UCL	1	1	1	0.92
Mixed (n = 5)	% Agreement	63.74	62.20	37.14	14.51
	Gwet AC1	0.40	0.53	0.20	0.04
	95% LCL	-0.12	0.18	-0.01	-0.02
	95% UCL	0.92	0.87	0.40	0.11
Sig. Error (n = 5)	% Agreement	75.38	67.25	40.44	21.54
	Gwet AC1	0.57	0.57	0.27	0.13
	95% LCL	-0.13	-0.03	-0.17	-0.10
	95% UCL	1	1	0.70	0.37
Poor (n = 4)	% Agreement	89.83	87.36	61.13	34.62
	Gwet AC1	0.89	0.85	0.52	0.25
	95% LCL	0.64	0.49	-0.05	-0.25
	95% UCL	1	1	1	0.76

It is also notable that inter-rater reliability varied between the 4 practical tasks (Table 4). Task 1 (setting up a burette) elicited the lowest inter-rater reliability while Task 2 (thin layer chromatography) elicited the highest reliability. This suggests that the tasks themselves introduced varying levels of ambiguity or challenge for the examiners. However, the confidence intervals were again too wide for strong conclusions to be drawn.

Table 4. *Reliability coefficients, grouped by task*

		2 Grades	3 Grades	5 point scale	Ten point scale
Task 1	% Agreement	58.24	52.97	31.43	13.85
	Gwet AC1	0.20	0.36	0.16	0.03
	95% LCL	-0.33	-0.04	0.02	-0.03
	95% UCL	0.73	0.77	0.30	0.09
Task 2	% Agreement	86.59	70.33	37.58	18.90
	Gwet AC1	0.75	0.59	0.22	0.10
	95% LCL	0.35	-0.06	0.04	0.02
	95% UCL	1	1	0.41	0.18
Task 3	% Agreement	78.46	66.81	46.59	26.37
	Gwet AC1	0.59	0.54	0.34	0.18
	95% LCL	-0.01	0.01	-0.18	-0.19
	95% UCL	1	1	0.86	0.55
Task 4	% Agreement	68.13	58.02	43.74	20.66
	Gwet AC1	0.41	0.43	0.31	0.12
	95% LCL	-0.25	-0.13	-0.11	-0.12
	95% UCL	1	1	0.72	0.36

Taken together, these findings indicate that, perhaps unsurprisingly, the nature of the task and the characteristics of the performance can have an impact on inter-rater reliability. This reflects the literature on marking reliability for written examination items (eg Black, Suto, & Bramley, 2011). A multiple linear regression was carried out to ascertain the extent to which task and performance type predict percentage agreement for the 2 grade rating scale. The model predicted 59.1% of the variance, though may not have been suitable for predicting the outcome (F = 2.479, df = 7, p = .08). The coefficients for the explanatory variables are shown in Table 5:

Table 5. *Regression coefficients for variables predicting inter-rater agreement.*

	B	SE	T	Sig.
Constant	.770	.105	7.352	.000
Task:				
Task 2	.239	.108	2.218	.047
Task 3	.202	.104	1.947	.075
Task 4	.071	.108	0.661	.521
Task 1 (reference)	0			
Type:				
Significant Error	-.167	.111	-1.503	.159
Mixed	-.250	.111	-2.250	.044
Creative	-.388	.148	-2.628	.022
Good	-.135	.116	-1.159	.269
Poor (reference)	0			

According to this model, in keeping with the inter-rater reliability coefficients reported earlier, ‘Mixed’ and ‘Creative’ performance types predict lower levels of reliability relative to the ‘Poor’ performance type. Relative to judgements for Task 1, Task 2 elicits higher levels of inter-rater agreement.

5 Discussion

5.1 Summary of findings

It would appear that binary (Pass/Fail) judgements regarding competence are not always straightforward and uncontroversial, even for seasoned subject experts. The results of this study provide modest evidence that, when holistically assessing proficiency in particular practical activities, examiner judgement is most reliable when differentiation is over 2 or 3 scale points. Larger numbers of scale points appear to be too granular for reliable holistic judgements. In addition, it is unsurprising to find some evidence that inter-rater agreement may be affected by both the nature of the performance that is observed (with poor performances easier to assess reliably than mixed ones) and the nature of the practical task itself.

It is also of note that examiners were divided between the 10-point scale and the 3 grade scale in terms of their preferences. Interestingly, those that preferred the longer scale seemed to do so because they believed it would allow them to provide more nuanced and encouraging formative feedback to students. Those who preferred the 3-point scale did so because it allowed them to be more confident in their judgements. Arguably, this may suggest that greater numbers of scale points may be preferable for formative assessment while fewer scale points may be preferable for summative assessment, at least from the perspective of teachers.

5.2 Limitations

This study was explorative in nature and the findings should be interpreted with care. There were 5 notable limitations:

1. Due to the generally low levels of agreement that were observed, the confidence intervals for the Gwet coefficients were very wide for the 2 and 3 grade rating scales.

2. As discussed in the methodology, the current study deliberately provided only limited guidance to examiners in the use of each rating scale. Examiners were asked to make their proficiency judgements holistically, without the use of a marking scheme or any other mechanism for mapping performances to scale points. Further guidance may have had a differential impact on the reliability of the 4 rating scales, possibly improving the relative reliability of the 5 and 10 point scales in particular. The findings of this paper do not therefore suggest that practical performance cannot be scored using a 5 or 10 point scale but rather that the ‘Fail/Pass’ or ‘Fail/Pass/Merit’ scales may prove a better fit for the holistic judgements that examiners naturally make during assessment. In

other words, using fewer grades may fit more readily with how examiners perceive proficiency and competence in practical work.

3. The language used to describe rating scales differed for the 2 and 3 grade scales in comparison to the 5 and 10 point scales, causing a confound in the research design. The 2 and 3 grade scales use the term 'competency' to direct examiners in their judgement while the 5 and 10 point scales use language that reflects more unrestrained judgements about the quality of the performance. This decision was made to avoid providing examiners with an artificial 'pass' mark for the scales with more gradation, but in hindsight it may have been better for all scales to have used the exact same terminology. None the less, as the next paragraph will suggest, examiners may have been ignoring the terminology and using all 4 grading scales with regard to the assessment of 'competency'.

4. The use of a repeated measures design meant that examiners were aware of all 4 rating scales as they were evaluating each performance. This meant that they may have been equating scale points across the rating scales, perhaps using a judgement made for one of the scales to anchor or inform their judgements for the others. For example, examiners may have decided that a 'pass' on the 3 grade scale was the equivalent to a '3 or more' on the 5 point scale. Though counterbalancing was employed to mitigate this effect (eg the order in which examiners were to apply scales was randomised), there was some evidence that it may have been taking place. Table 6 shows the relationship between these 2 scales to demonstrate the appearance of such conceptual 'grade boundaries'. Arguably, this threshold may represent a level of proficiency at which the examiner feels the student has demonstrated competency.

Table 6. Cross-tabulation and chi-square for 3 and 5 point scales

		5 point scale				
		1	2	3	4	5
3 grades	Fail	66	79	22	3	0
	Pass	0	1	40	37	4
	Merit	0	0	0	10	18

$\chi^2(8, N = 280) = 321.76, p < .001$

5. The use of pre-recorded video footage of practical work mean that examiners in the study could independently assess identical performances. However, video is an imperfect proxy for live observation. The camera dictates where the examiner focuses their attention and may fail to capture details which the examiner would consider important. Live

observation allows the examiner to move around and focus wherever they choose. The advantage of pre-recorded footage is that it allows the examiner to repeatedly review the performance, or to scrutinise particular aspects of it.

5.3 Considerations for the A level science practical endorsement

These findings relate only to assessing the performance of particular practical techniques and do not generalise to assessing the CPAC for the A level practical endorsement. For the CPAC, the intention is to evaluate a broader conception of practical skills and to assess performance over the duration of a 2-year course. CPAC judgements are therefore more holistic and general than judgements made about the performance of specific practical activities at a particular time.

This is not to suggest that criteria like the CPAC could not be assessed over more than two grades, just that this research does not provide evidence either way. What does seem apparent from this study is that defining the standard of those grades and ensuring their consistent assessment would be a complicated matter. Exam boards would need to work with Ofqual to decide which criteria warranted an additional grade and to develop a grading system that was unambiguous, fair and logistically accessible for all schools and colleges.

Any change to the grading system would also have to be considered with regard to possible unintended consequences. If additional grades were available, there would likely be pressure on teachers to ensure that their students were achieving the best possible outcomes. Such pressure may inadvertently encourage an overly prescriptive approach to teaching practical work or difficulties in policing malpractice and ensuring fairness. Changing the current grading structure for the CPAC would also impose significant burden on awarding organisations and teachers, a consideration which is not trivial given the intensity of recent qualification reform. There would need to be compelling evidence for action.

It is important to note that refinements to the current arrangements can be made without substantive systemic change. For example, one of the subsidiary findings from this study implies that some teachers may prefer to assess practical work using more than 2 grades so that they can be more nuanced in their feedback to students. The new assessment arrangements are in part designed to empower teachers to deliver practical work in the way that best suits their students. Any positive washback effects that the practical endorsement has on teaching and learning is therefore likely to be dependent on the nature of the formative feedback that is provided for students as they work towards the CPAC.

There is nothing in the current regulatory requirements that prevent teachers from providing formative feedback in whichever way they feel most appropriate, as long as

they can provide evidence to exam boards that the student has met the required standard across the CPAC by the end of their course. Exam boards and other stakeholders will continue to play a supportive role in this regard, by providing teachers with advice and materials..

5.4 Conclusion

This study took an innovative approach to comparing examiner ratings of performance and raises some interesting issues. The findings suggest that an individual's proficiency in the performance of particular practical techniques can be discriminated beyond binary judgements of competence, but that such judgements are complex and would require significant guidance and standardisation in order to achieve acceptable levels of inter-rater reliability across the full range of possible tasks and performances. It would appear that deciding whether or not the performance of a given technique is 'competent' can often be quite a nuanced and challenging task.

6 References

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. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/03057267.2013.858496

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. http://doi.org/10.1080/09500690902777410

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 259–278. http://doi.org/10.1080/0969594X.2010.546775

Black, B., Suto, I., & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 295–318. http://doi.org/10.1080/0969594X.2011.555328

Blood, E., & Spratt, K. (2007). Disagreement on Agreement: Two Alternative Agreement Coefficients. In *SAS Global Forum*. Retrieved from http://www2.sas.com/proceedings/forum2007/186-2007.pdf

Department for Education. (2014). GCE AS and A level subject content for biology, chemistry, physics and psychology. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/593849/Science_AS_and_level_formatted.pdf

Gatsby. (2012). *Science for the Workplace*. Retrieved from http://www.gatsby.org.uk/education/reports?filter[]=310

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

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. http://doi.org/10.1080/03057640220147540

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

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*,

61(1), 29–48. http://doi.org/10.1348/000711006X126600

Harlen, W. (1999). Purposes and Procedures for Assessing Science Process Skills. *Assessment in Education: Principles, Policy & Practice*, *6*(1), 129–144. http://doi.org/10.1080/09695949993044

Lunetta, V. N., Hofstein, A., & Clough, M. P. (2007). Learning and Teaching in the School Science Laboratory: An analysis of research, theory, and practice. In S. Abell & N. Lederman (Eds.), *Handbook of Research on Science Education* (pp. 393–431). Mahwah, NJ: Lawrence Erlbaum Associates. Retrieved from https://www.researchgate.net/publication/272680843_Learning_and_teaching_in_the_school_science_laboratory_An_analysis_of_research_theory_and_practice

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. http://doi.org/10.1080/0969594X.2010.496173

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. http://doi.org/10.1080/09695940701478321

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. (2017). *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

Pinot de Moira, A. (2011). *Levels-based mark schemes and marking bias*. Manchester, UK: AQA Centre for Education Research and Policy.

Pinot de Moira, A. (2013). Features of a levels-based mark scheme and their effect

on marking reliability. Manchester: AQA Centre for Education Research and Practice. Retrieved from https://cerp.aqa.org.uk/research-library/features-levels-based-mark-scheme-effect-marking-reliability/how-to-cite

Pye Tait. (2014). *Competence in Construction*. Retrieved from http://www.citb.co.uk/news-events/report-highlights-need-for-construction-competency-framework/

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

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.

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, *13*(1), 61. http://doi.org/10.1186/1471-2288-13-61

7 Annex A: Pre- and post-reform assessment of practical skills at A level

Reformed A level science qualifications were introduced for first teaching in September 2015 and the assessment arrangements for practical work changed significantly. The pre- and post-reform approaches can be described as follows:

- For *pre-reform* specifications, practical work was assessed via 'Non Examined Assessment' (NEA) components. These components contributed to a student's final grade (with a weighting of 20%) and required students to complete a practical activity (or activities) under controlled conditions. The NEA could take a variety of forms and was either 'externally' marked by exam boards or 'internally' marked by teachers (and externally moderated by exam boards). The nature of the NEA components varied between exam boards and specifications, with some requiring students to complete an individual investigation over a period of weeks and others requiring students to complete one or more scaffolded practical tasks within a specific time frame. Only a small percentage of the total marks for the NEA components were allocated to the direct observation of practical work, with the majority of marks allocated to written work (eg planning and data analysis).

- For *post-reform* specifications, assessment of practical skills is achieved in 2 ways (see Ofqual, 2015). Firstly, 15% of all available marks in the written examinations are allocated to questions which indirectly assess practical skills. Secondly, students must be given opportunity to demonstrate their competency in a range of practical techniques and using a range of apparatus, completing at least 12 'hands on' practical assignments. Throughout their course, students are assessed (by their teacher) against 5 Common Practical Assessment Criteria (CPAC)[15] and receive one of 2 grades: either 'Pass' or 'Not Classified'. This element of the assessment does not contribute to a student's primary grade, instead they receive a second, separate grade (an 'endorsement') when they certificate. Schools and colleges are visited by an exam board 'monitor', whose role is to ensure that students are provided with appropriate opportunity

[15] The 5 Common Practical Assessment Criteria (CPAC) are: follows written procedures; applies investigative approaches and methods when using instruments and equipment; safely uses a range of practical equipment and materials; makes and records observations; researches, references and reports (see Ofqual, 2016, pp. 15–16 for further details).

to undertake practical work and that adequate records of activities and achievements are being maintained for the endorsement.[16]

8 Annex B: 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:

- Paper 1: Teacher interviews – Perspectives on A level reform after one year
- 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

This study (Study 3) is somewhat separate to the others in the research programme in that it seeks to explore one of the issues associated with the direct assessment of performance in the context of practical work, something that is a longstanding challenge for many qualifications, not just science.

[16] Any school or college that offers an A level in science must receive a monitoring visit from an exam board at least every 2 years (Ofqual, 2016).

9 Annex C: Practical tasks & performance indicators

Chemistry practical skills tasks

Outlines and performance indicators

Practical Task 1: Setting up a Burette

Instructions to candidates:

You should set up a burette containing 0.1 mol dm^{-3} HCl as if you were titrating it against an alkali. You will be expected to set the initial volume to a value between 0.00 cm^3 and 10.00 cm^3. You may not need all of the apparatus that is provided and should select what you do need.

Equipment available:

- Burette
- Suitable clamp
- Filter funnel
- Filter paper
- White tile
- Distilled/deionised water in wash bottle
- 0.1 mol dm^{-3} HCl
- 25 cm^3 volumetric pipette and pipette filler
- 250 cm^3 conical flask

- 250 cm^3 beaker

Performance indicators:

We would like you to evaluate the performance of the candidate as they complete the task above. We would like you to apply your professional judgement and consider any aspects of the task which you feel are important, but you may wish to consider the following:

- Does the candidate select the most appropriate equipment?
- Does the candidate handle the equipment and materials safely?
- Does the candidate perform the techniques correctly?
- Does the candidate take accurate readings (eg to nearest 0.05 cm^3 of initial level of solution)?

You may also wish to look out for the following examples of good practice:

- Rinses burette with HCl, or rinses with distilled H_2O and then HCl
- Clamps burette vertically
- Fills burette safely below eye level (funnel not essential, but if funnel is used then it must be removed after filling)
- Opens tap to ensure that jet is full of acid (no air bubbles)

Practical Task 2: Thin Layer Chromatography

Instructions to candidates:

Aspirin can be produced from salicylic acid. The progress of the reaction can be monitored by removing samples from the reaction mixture during the procedure and performing Thin Layer Chromatography (TLC). You should set up a TLC plate to verify the whether the 'Reaction Sample' contains pure aspirin, pure salicylic acid or a mixture of both. You should use a 50:50 % ethanol/dichloromethane solvent to dissolve the aspirin and ethylethanoate as the mobile phase. There may not be sufficient time for you to analyse your results fully but you are expected to get the TLC plate running. You may not need all of the apparatus that is provided and should select what you do need.

Equipment available:

- TLC plates
- Capillary tube or similar

- Ruler
- Pen and pencil
- Filter paper
- Solvents eg ethanol, dichloromethane, ethylethanoate
- Distilled/deionised water in wash bottle
- Samples of salicylic acid, aspirin and 'Reaction mixture'
- Pasteur pipettes
- Suitable TLC Tank, holders and lid
- 100 cm^3 beakers
- Measuring cylinders
- Tweezers
- Hairdryer

Performance indicators:

We would like you to evaluate the performance of the candidate as they complete the task above. We would like you to apply your professional judgement and consider any aspects of the task which you feel are important, but you may wish to consider the following:

- Does the candidate select the most appropriate equipment?
- Does the candidate handle the equipment and materials safely?
- Does the candidate perform the techniques correctly?

You may also wish to look out for the following examples of good practice:

- Ensures sample is securely held and is not below the solvent
- Draws baseline about 1-2 cm from one end of plate using a pencil
- Doesn't touch TLC plate surface with fingers
- Measures volumes of solvents to ensure appropriate 50:50% composition to dissolve the 3 samples (~5-10 cm^3)
- Applies samples to plate using capillary tube – small 1-2 mm, well-spaced spots, may use pencil line to ensure same distance

- Adds a volume of solvent to the tank to ensure that the solvent is **below** the level of the dots on the plate.

Practical Task 3: Setting up a reflux and distillation

Instructions to candidates:

Part 1: Ethanol can be oxidised to ethanoic acid by refluxing with a suitable oxidising agent. You are required to set up the equipment needed for refluxing aqueous reactants. You are not required to add any chemicals or to actually heat the apparatus. You may not need all of the apparatus that is provided and should select what you do need.

Part 2: After the alcohol has been oxidised for an appropriate period of time, the ethanoic acid can be distilled from the reaction mixture. You should now reconfigure your apparatus so that it can be used to distil off the organic product.

Equipment available:

- Access to cold tap and nearby sink/drain
- Pear shaped flask
- Round bottomed flask
- 400 cm^3 beaker
- Heating mantle
- Bunsen burner
- Tripod and gauze
- Leibig condenser
- Glass stopper
- Quickfit thermometer
- Quickfit adaptor
- Quickfit delivery tube
- Quickfit clips
- Anti-bumping granules
- 3 clamps, 3 bosses, 2 retort stands

Performance indicators:

We would like you to evaluate the performance of the candidate as they complete the task above. We would like you to apply your professional judgement and consider any aspects of the task which you feel are important, but you may wish to consider the following:

- Does the candidate select the most appropriate equipment?
- Does the candidate handle the equipment and materials safely?
- Does the candidate perform the techniques correctly?

You may also wish to look out for the following examples of good practice:

- Use of anti-bumping granules in flask
- If pear shaped flask is used, heating should be provided by 400 cm^3 beaker as water bath, on gauze/tripod with Bunsen underneath
- If round bottomed flask is used, heating should be provided by heating mantle
- Condenser is fitted vertically into flask and is open (ie it does not contain a glass stopper)
- Water from tap enters bottom of condenser and leaves from the top to a suitable drain/sink
- All pieces of apparatus are effectively secured (using clamp(s) and possibly quick fit clips)
- Uses adaptor to mount condenser on a decline for distillation
- Fits delivery tube to condenser
- Fits thermometer into top of distillation adaptor with thermometer bulb at level of condenser branch in adaptor

Practical Task 4: Making up a standard solution

Instructions to candidates:

You are required to make up a standard solution of sodium carbonate, with a concentration of 0.100 mol dm^{-3}. This is achieved by dissolving 2.65 g of anhydrous sodium carbonate and making this into 250 cm^3 of solution. You may not need all of the apparatus that is provided and should select what you do need.

Equipment available:

- 250 cm^3 volumetric flask and stopper

- 250 cm^3 beaker
- 100 cm^3 beaker
- Plastic or glass dropper pipette
- Distilled/deionised water
- Filter funnel
- Anhydrous sodium carbonate, approximately 4 g Spatula
- Top pan balance
- 2 x Weighing boats
- Glass rod

Performance indicators:

We would like you to evaluate the performance of the candidate as they completed the task above. We would like you to apply your professional judgement and consider any aspects of the task which you feel are important, but you may wish to consider the following:

- Does the candidate select the most appropriate equipment?
- Does the candidate handle the equipment and materials safely?
- Does the candidate perform the techniques correctly?
- Does the candidate take accurate measurements (eg to the nearest 0.01g of solid)?

You may also wish to look out for the following examples of good practice:

- Zeroes balance after putting weighing boat or beaker on
- Does not spill any solid on balance or bench
- Correctly transfers all solid to flask. This could be done either by adding solid through funnel and rinsing funnel into flask, or by dissolving the solid with some water in the beaker and then transferring the solution into the flask through the funnel and rinsing the beaker and funnel into the flask.
- Swirls the flask to dissolve the solid completely (before filling the neck of the flask)
- Adds water so that the bottom of the meniscus is on the mark on the neck of the flask (probably using dropper pipette)
- Inserts stopper and inverts a suitable number of times

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

© Crown copyright 2018

This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. To view this licence, visit http://nationalarchives.gov.uk/doc/open-government-licence/version/3 or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU, or email: publications@ofqual.gov.uk.

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned.

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