The assessment of chemistry subject knowledge in secondary education: a critical evaluation of the literature

Final Report to the Royal Society of Chemistry, April 2017

Judith Bennett
Lynda Dunlop
Kerry J. Knox
Mary Whitehouse

UNIVERSITY OF YORK
SCIENCE EDUCATION GROUP

Review team

Judith Bennett, University of York

Lynda Dunlop, University of York

Kerry J. Knox, University of York

Mary Whitehouse, University of York

Contact details

Professor Judith Bennett

Department of Education

University of York, UK

York YO10 5DD

Tel: 01904 323471

email: judith.bennett@york.ac.uk

Comments on the report

Every effort has been made to ensure that the contents of this report are accurate. If you have any comments on matters of factual accuracy, we would be happy to discuss these with you. Please contact the review team leader, Judith Bennett.

Acknowledgements

UYSEG thanks the key informants for their time and their contribution to this review.

Citation

This report should be cited as: Bennett, J., Dunlop, L., Knox, K. J., and Whitehouse, M. (2017) *The assessment of chemistry subject knowledge in secondary education: a critical evaluation of the literature: Final report to the Royal Society of Chemistry, April 2017*. York: Department of Education, University of York.

Contents

List of tables

Executive summary

1. Assessment of subject knowledge in chemistry performs an important and necessary role as it provides information about pupils' progress and achievements. This review focuses on summative assessment, i.e. assessment that measures and reports on learning outcomes in order to report and make comparisons.

The focus of the review

2. This review comprised two main strands: a systematic exploration of the published and grey literature on the summative assessment of chemistry subject knowledge for young people aged 14-19 since 2006, and interviews with 14 key informants with expertise in the assessment of chemistry in the UK and internationally.

The review methods

Literature review

3. The literature review was conducted in accordance with the procedures normally associated with systematic review.

4. Four strategies were used to identify relevant literature: electronic searches of standard databases; recommendations made by key informants during interviews; hand searches of recent journals; and literature already known to the research team.

5. Publications were included in the review if they met the criteria drawn up for the review. The overall approach was inclusive, i.e. where publications offered something of relevance for the review they were included, even if insufficient details were included to enable all the inclusion criteria to be met.

6. The detailed review is based on 33 publications that reported on models for summative assessment relevant to the assessment of chemistry subject knowledge.

The interviews with key informants, including examiners

7. Interviews were conducted with 14 key informants with expertise in the assessment of chemistry subject knowledge. These included examiners from British and international awarding bodies and qualification authorities, and academics with expertise in chemistry assessment. These included individuals with experience in international contexts including Australia, Germany, Israel, Singapore and the USA. The interviews were used to identify practical considerations for different assessment models, and to suggest future directions in the assessment of chemistry subject knowledge.

Evidence from the review

8. The review found evidence published since 2006 on the following assessment instruments: multiple choice questions, closed response questions, open response questions, and performance assessments. Most assessment systems use a combination of different types of instrument. The review also found evidence relating to different scoring systems: points-based and levels-based mark schemes, and comparative and adaptive comparative judgement.

9. Almost all the literature used in this review was published in peer-reviewed journals.

The nature of the research

10. The research on the assessment of chemistry subject knowledge could be considered to fall into four broad categories:

    a. research concerned with developing diagnostic assessments to develop the understanding of students' conceptions of chemical ideas

    b. research which focuses upon a particular model of assessment, and it may be coincidental that the subject being assessed is chemistry; this research is often about reliability of systems of assessment rather than about the content validity of the assessment

    c. research aimed at developing summative assessment tools that can effectively evidence students' understanding of chemistry, for example work aimed at developing MCQs that assess learning beyond factual recall, or the development of new forms of assessment

    d. small-scale action research projects undertaken in a small number of institutions (often a single institution) and aimed at assessing the effectiveness of a particular form of assessment such as the use of oral examinations in chemistry; this research generally takes place in undergraduate programmes.

Multiple choice questions

11. Multiple choice questions (MCQs) are a common example of fixed response questions. These consist of a question stem and a choice of answers. These are used widely, particularly in the USA where they lend themselves to efficient assessment of large cohorts. Recent research on the use of MCQs for the assessment in chemistry has included that focused on the development of questions that require higher level thinking skills, questions to assess students' levels of understanding, and the use of polytomous scoring systems.

12. Evidence from the literature and from key informants identified MCQs as reliable in the context of the assessment of chemistry, allowing assessment of the full specification, and rapid marking (often by machine). They are, however, often perceived as 'easy'; the possibility for students to answer correctly by chance or educated guessing was noted. These items can be difficult to write, and it can be challenging to recruit question setters with the necessary expertise. Future directions in the use of on-screen assessments may lead to the use of more 'two-tier' MCQ, and the partnering of MCQs with embedded simulations or extended explanations.

Closed response questions

13. Closed response questions are useful for assessing recall of chemistry ideas and students' ability to apply a chemical idea in a different context from that in which it was taught. This type of assessment consists of objective questions to which the students provide an answer, constrained by the format of the question. Examples include the labelling of diagrams, single word answers and solutions to numerical problems. These items can be scored reliably and have the potential for on-screen assessment with computer scoring.

14. In common with fixed response questions, closed response questions are suitable for assessing recall and application of knowledge but do not lend themselves well to assessing some higher order thinking skills.

Open response questions

15. Open (or constructed or free) response questions are those that do not constrain students' responses. Responses may be written, drawn, or calculated (or a combination of these), and are usually marked by examiners. They include short answer questions, concept maps, and

extended written responses such as essays. Concerns exist over the reliability of marking open response questions.

16. Although there is good correlation between students' scores on short answer questions (SAQs) and MCQs when assessing chemistry, these do not measure the same constructs; SAQs have been found to be more difficult than MCQs, and in some studies females have been found to score higher on tests including SAQs rather than those containing only MCQs.

17. Concept maps are used to help students make connections between concepts. It is difficult to create a consistent model for scoring students' constructed concept maps because of the variety of ways in which students can present their ideas. However, 'creative exercises,' in which students identify as many distinct, correct and relevant statements in response to a prompt show some signs of promise. Research in the US suggests good inter-rater reliability and moderate correlation with performance on MCQ tests.

18. Open book examinations allow students to use reference materials during an examination. The types of open book examinations used for the assessment of chemistry subject knowledge have used open response questions. Studies suggest that this approach allows students to develop a better understanding of the process of learning, the nature of knowledge, and higher order thinking skills. There is a need to prepare students for this approach to assessment; one study found those students given an open book examination performed more poorly than those given the same assessment as a closed examination.

19. Open response questions can be scored using points-based mark schemes (used for most open response questions in chemistry) or levels-based mark schemes. The latter are used for longer prose answers and include descriptions of criteria that a student's response must match to be awarded marks in a certain 'band'. Criterion-referenced systems allow differentiation by outcome, and are transparent to students and teachers.

20. A number of research studies have developed frameworks for developing levels-based mark schemes and associated tasks. These include the development of criteria for application of content knowledge, interconnectedness between ideas, and for explanations and argumentation.

Performance assessments

21. Practical work and extended projects are often assessed by the teacher, followed by a moderation process. These are examples of performance assessments, which seldom have the main purpose of assessing chemistry subject knowledge. Although both practical work and extended projects require conceptual understanding, the focus of the assessments is on skills that cannot be assessed in a closed examination paper.

22. Key informants generally reported that performance assessments were valid assessments that encourage students to think and act like a scientist, but noted that they were sometimes perceived as being open to interference, presenting a role conflict for teachers, and requiring of high demands from teachers and technicians including time, workload, and space.

23. Other performance assessments of chemical subject knowledge and understanding currently in use include oral examinations and web-based video responses, both of which have been reported for use with undergraduates. Oral examinations were found to promote creativity and a high level of learning, and to be efficient in terms of staff time. Video resources created by students were perceived to help students improve their chemical knowledge and to allow rapid feedback from tutor to student, but required increased staff time. Video responses are already used in some other subjects in secondary education.

The depth and quality of the evidence base

24. The depth and quality of the research evidence base is variable. Research into the refinement of particular approaches (e.g. MCQs, computer assisted scoring) appears to be generating a solid evidence base. Research into the use of new assessment is often characterised by work being undertaken by advocates for a particular method. Where the work involves only one (or two) institutions the evidence base is slight.

Future directions

25. The possibilities provided by modern computer systems are likely to influence the assessment of chemistry subject knowledge. Examples of likely developments include the use of computers to mark work and the use of varied stimuli as the basis for assessment items. These stimuli are expected to include simulations, visualisations, virtual laboratory contexts, digital portfolios, game-based approaches, and virtual or augmented reality situations.

26. Computer-assisted scoring of open response items is a growing area of research interest. The driver here is to develop a means to assess students' use of evidence and argument construction, neither of which lend themselves well to assessment by MCQ. This involves 'training' a computer using expert human raters. Recent research has found computer marking to be as accurate and reliable as human marking, with a number of caveats: spelling and linguistic variation can result in differences between human and computer marking, and between 500 and 1000 responses need to be marked by hand to build a reliable model.

27. Comparative judgement and adaptive comparative judgement allow teachers' professional judgement to replace traditional 'marking' and are a response to concerns over reliability of marking open response questions. This involves a person deciding which of two responses is better, rather than allocating a score. Many such judgements are made, resulting in the responses being placed on a point on a scale. Research on the use of comparative and adaptive comparative judgements indicates that these are valid and reliable forms of assessment. Researchers have also used the model successfully for peer assessment. Little training is required other than how to use the software, but, there are concerns about how acceptable these methods will be for high stakes assessments.

28. Computer-based approaches that require students to complete assessment tasks on screen will require widespread access to the appropriate technology. Infrastructure and security concerns are potential barriers to this.

29. Although the scope of the review did not seek to understand the role of regulation in the assessment of chemistry subject knowledge, a number of key informants identified challenges associated with regulation. These challenges included the pace of change in regulations and the limited scope for innovation when there are stringent regulations over what is assessed and how.

Further work

30. Further research could focus on the disciplinary dimensions of assessing chemistry, for example, the challenges for students associated with switching between microscopic, macroscopic, and symbolic representations during assessments, and the use of models in assessments, particularly the extent to which a task assesses content knowledge or understanding of the model.

31. Further work could also focus on assessment of chemistry subject knowledge through vocational qualifications, extension assessments such as the Chemistry Olympiad and Cambridge Chemistry Challenge, university admissions assessments and on the online and on-screen assessment experiences of the Open University.

32. In-depth case studies of international assessment systems could be used to inform the future direction of chemistry assessment in the UK and Ireland. For example, the development, adoption, and assessment of the Next Generation Science Standards in the USA could inform advances in the use of MCQs and electronic assessment to assess chemistry subject knowledge. Likewise, research and development associated with international assessments such as the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) could provide relevant insights.

Section 1: Introduction

1.1 Aims of the review

This review of the summative assessment of chemistry subject knowledge in secondary education addresses three questions:

1. What are the main summative assessment models for the assessment of chemistry subject knowledge in secondary education?
2. What are the advantages and disadvantages of each model that are reported in the literature?
3. What is the depth of research understanding available about the effectiveness of each assessment model?

Within these questions, the review has sought to gather information on:

- assessment models that have some subject knowledge element, including integrated assessment models (e.g. approaches that assess both subject knowledge and skills)
- assessment models that could be used to assess chemistry but where the literature available doesn't refer to chemistry applications
- past as well as current assessment models
- international work and perspectives on assessment of chemistry subject knowledge.

The review comprises two main strands of activity. The first strand is a systematic exploration of published and grey literature covering the summative assessment of chemistry subject knowledge in the 15-19 age range. The second strand consists of interviews conducted with 14 key informants with particular perspectives on assessment in chemistry. These two sources of data have been combined into a synthesis of approaches to assessment and, where applicable, research into their effectiveness.

1.2 Context of the review

Assessment of pupils' performance in science plays an important and necessary role. It provides information about pupils' progress and achievements – information which should be of interest to the pupils themselves, to their teachers and to their parents. In the words of Black (1990):

> Assessment is at the heart of the process of promoting children's learning. It can provide a framework in which educational objectives may be set and pupils' progress charted and expressed. It can provide a basis for planning the next educational steps in response to children's needs. By facilitating dialogue between teachers, it can enhance professional skills and help the school as a whole to strengthen learning across the curriculum and throughout its age range (p27).

However, assessment has been very much at the forefront of debate in the last two decades or so as it has become increasingly associated with accountability, the drive to raise standards, and the production of league tables of pupil performance in national and international tests. Moreover, questions continue to be asked about the form assessment might take, what should be assessed, the extent to which teachers should be involved, how often assessment should take place, and what should be done with data on pupils' performance.

An effective assessment system needs to give careful consideration to the aspects of a curriculum that are going to be assessed, the ways in which the assessment should be carried out, how the results will be interpreted, and the uses that will be made of the results.

What role should assessment play in the teaching of chemistry and science? A desirable assessment system would draw on a range of methods to yield sufficient reliable and valid data to enable standards to be monitored whilst, at the same time, provide appropriate measures of pupils' understanding – with the latter, ideally, permitting feedback to be provided to students to assist with their learning.

This review has considered a diversity of assessment techniques relevant to the teaching of chemistry. The techniques lie in the area of the *summative* assessment, i.e. assessment that is normally formal in nature, that takes place at the end of a teaching block, and that aims to measure and report on learning outcomes in order to make a variety of comparisons.

1.3 The review report

The review report has seven main sections. Section 2 provides details of the review methods. Section 3 provides an overview of the literature used in the report. Sections 4 and 5 consider, in turn, the variety of assessment tasks presented in examinations and the ways in which performance assessments are used in chemistry education. Section 6 considers future directions in the assessment of chemistry in secondary education. Section 7 summarises the evidence and proposes some points for further consideration.

Section 2: Review methods

2.1 Identifying the relevant literature and research studies

Four strategies were employed to identify the relevant literature.

1. Searches were carried out of the standard electronic databases available: the Education Resources Information Centre (ERIC), the British Education Index (BEI), the Social Sciences Citation Index (SSCI) and PsychINFO. The search focused on the period 2006 to 2016.
2. Key informants were asked to identify relevant publications during their interviews.
3. Hand-searches of journals were carried out to identify any very recent publications that may not yet have been listed on electronic databases.
4. In addition, the research team added a small number of other publications of which its members were aware and felt were relevant to the review. This included the identification of a small number of pre-2006 publications which provide an excellent overview of the issues surrounding summative assessment.

Full details of the electronic search strategy may be found in Appendix 1.

There were a number of challenges associated with the review.

Many publications about assessment focus on the use of assessment to evaluate an intervention, to develop diagnostic assessments, or formative assessment; these publications often do not provide information that would be useful when considering summative assessment.

Publications on assessment often include those in the 'grey' literature. Grey literature is often less accessible than academic science education literature as, in the context of summative assessment, it consists of, for example, reports from groups involved with the setting of examinations.

The websites of the three awarding bodies in England were used to identify details of current assessments of chemistry at GCSE and Advanced level.

The initial electronic searches identified 1148 publications. This was reduced to 256 after publications describing only diagnostic assessments, formative assessment methods, or evaluation of interventions, were removed. Further refining using the inclusion criteria described in Appendix 2 reduced the number of publications used in the review to 30. The key informants, hand-searches and publications identified by the research team added a further 3 publications to the main review.

2.2 Defining relevant studies: inclusion criteria

In order to identify the relevant literature, a number of inclusion and exclusion criteria were developed for the work reported in the literature. These were informed by the review research questions.

The inclusion criteria may be found in Appendix 2.

The overall approach was to be inclusive, i.e. where publications appeared to offer something of relevance to the review; they have been included, even if there were insufficient details to enable all the inclusion criteria to be applied.

Application of the inclusion criteria yielded 33 publications which formed the basis of the report. A number of other publications provided useful general background material or overviews of provision, without reporting any data in detail. These publications have also been included in this report.

2.3 Extracting the key information from the literature

A bespoke data extraction sheet was developed for extracting the key information from the publications. This focused on the following information:

- practical details (author, title, year of publication, source, country of origin, details of the researchers)
- the type of assessment methods described in the publication and the name of the assessment programme, where applicable
- the subject(s) being assessed
- the age(s) of the learners being assessed
- a brief description of the assessment method
- reported or apparent advantages and disadvantages of the assessment method
- any other information worth noting
- for a publication including research, the following additional information was noted:
- the review research questions addressed by the publication
- aims of the study being reported
- study design, including details of the sample
- data collection methods and instruments (including reliability and validity checks)
- data analysis methods (including reliability and validity checks)
- summary of findings and conclusions
- any other information worth noting.

The two members of the research team undertaking the literature review (Bennett and Whitehouse) worked closely together on the development of the data extraction sheet, testing a pilot version on a small number of the publications and then fine-tuning the sheet to ensure it covered the key information needed.

A copy of the data extraction sheet may be found in Appendix 3.

2.4 The interviews with key informants

Interviews were conducted with 14 key informants with useful perspectives on assessment. The informants included those with experience of research in the area of assessment, and examiners for awarding bodies in the United Kingdom, including those with leadership responsibilities. Nine of the informants were currently based in the United Kingdom, while five were based abroad in Asia, Australasia, Europe, and North America; the perspectives offered by the informants tended to draw principally upon, but are not restricted to, assessment in the area of the world in which they currently work. The key informants work in countries where a substantial amount of work is being undertaken on assessment or where students perform well in international comparisons.

The report does not include direct quotations from key informants or the specific examples of assessments they discussed as much of this detail would be identifying. All key informants spoke on the basis that their contributions would be anonymous.

Section 3: Overview of the sources of evidence

3.1 Introduction

There is a large body of literature related to assessment, including the assessment of students' understanding of chemistry, however much of that assessment is for the purpose of evaluating an intervention, or to develop diagnostic assessments that will be used formatively. There is much less research on the effectiveness of assessment instruments for the summative assessment of chemistry.

This review focuses on identifying the assessment instruments that are being used, or have been used to assess chemistry subject knowledge in the 14-19 age group. The review includes a number of reports of work in universities where instructors are seeking to find more effective ways of assessing their students' understanding; these have been included where they may have the potential to be useful in a school setting.

3.2 The literature search

The literature search identified 64 publications from 9 countries that reported research into the assessment of chemistry and met the inclusion criteria for the review. On closer scrutiny, 31 of these publications did not provide sufficient information about the assessment tools or their efficacy for summative assessment. The review therefore covers the 33 publications from 7 countries that contribute information relevant to the assessment of chemistry subject knowledge for summative purposes.

32 out of the 33 publications appeared in peer-reviewed journals.

Countries in which the research included in this review was conducted

Research into the development of assessment approaches is undertaken in a number of countries across the world; those countries that feature in this review are shown in Table 1.

Table 1 Country of study

Country	Publications
Australia	3
Canada	3
Germany	3
Greece	1
Norway	1
UK	7
USA	15
TOTAL	33

Student age range

The review focused on recent research into the assessment of chemistry for the upper secondary/high school age range, taken to be ages 14-19. As Table 2 indicates, the research in the publications was most often focused on assessment at the undergraduate level in universities.

Table 2 Student age range identified in the research included in this review

Age of students	Publications
Middle high school (age 14-16)	4
High school (ages 14-19)	6
Senior high school (age 16-19)	7
University undergraduate	15
Age-independent	1
TOTAL	33

Science discipline

Most of the literature reviewed described the assessment of chemistry subject knowledge; however the report does include some evidence from other subjects where it was felt to be relevant and offered a different perspective. Table 3 shows the subjects that were the focus of the publications reviewed here.

Table 3 Disciplinary focus of the publications included in the review

Discipline	Publications
Biology	4
Chemistry	17
Physics	1
Engineering	1
Geography	1
Mathematics	3
Science	4
Discipline-independent	2
TOTAL	33

Types of assessment

The literature reviewed described a variety of ways of assessing science subject knowledge, including approaches to marking. Table 4 shows the number of publications related to each of the key approaches discussed in the review.

Table 4 Assessment approaches

Assessment approaches	Publications
multiple choice	9
closed response	1
open response	
open book	4
concept maps	2
short answer questions	2
pre-release materials	1
criterion-referenced marking	3
performance assessment	
practical work	4
oral examination	1
use of video	1
comparative judgement	4
computer-assisted scoring	2
fieldwork	1

Section 4: Assessment: question formats and scoring

This section of the report describes the types of questions that are used to assess chemistry subject knowledge through written papers taken under examination conditions.

4.1 Multiple choice questions

Multiple choice questions (MCQs) are a specific example of questions that require students to select an answer, rather than offer an answer of their own construction; such questions are described as 'fixed response questions'. Other models of fixed response questions include cloze tasks where students are asked to select words to complete statements, and matching questions, for example where students match properties of materials to particular substances.

This section of the report focuses on multiple choice questions, which is where much of the development is taking place.

A multiple choice question consists of two parts:

- the stem in which the question is posed, and which also may contain stimulus information needed to answer the question
- choices of answers – the correct response, sometimes called the key or attractor, and also the incorrect choices, known as distractors.

MCQs are used to assess students' knowledge and understanding across a wide range of subjects. They have been used in the USA since the 1920s where large enrolments mean easily scored multiple choice tests are particularly useful (Black, 1998).

MCQs can take a variety of formats, for example the number of distractors provided can vary, or a piece of common information may be used to pose several questions. Typically MCQs are scored dichotomously – one mark for the correct answer and zero otherwise. The development of computer scoring has made it easier to use polytomous scoring systems (allowing scores other than 1 or 0) where partial credit can be given for selecting certain distractors.

It has been argued that MCQ can only be used to assess recall. However, the literature shows that provided MCQs are written appropriately, it is possible to test conceptual understanding and to learn a certain amount about students' higher level thinking skills. Likewise, some of the key informants argued that it was possible to write questions that challenge the best candidates.

4.1.1 Developments in the design of MCQs

In recent years a number of testing regimes have developed more complex questions that test higher level thinking skills, for example Domyancich (2014) describes how the MCQs that form part of the Advanced Placement tests in chemistry in the USA were redesigned to better test conceptual understanding and higher order cognitive skills by requiring students to apply concepts in unfamiliar contexts. The motivation for this was to better match the assessments to the intended curriculum.

There are groups looking at the development of MCQs across a wide range of subjects, this section reports on some recent developments of MCQs which aim to improve the discrimination and reliability in assessment of chemistry and, in one case, physics.

Assigning partial credit for partial knowledge

When constructing MCQs the examiner will typically attempt to base the distractors on the sorts of answers students might give to a free response item: in a question requiring a calculation the distractors may be alternative results of the calculation when a common error has been made; in questions testing understanding the distractors may be partially correct, or may reflect common misunderstandings.

The American Chemical Society Examinations Institute has explored the possibility of giving partial credit for students selecting incorrect answers that show partial knowledge in their multiple choice examinations (polytomous scoring) (Grunert, Raker, Murphy, & Holme, 2013). The researchers

worked with 23 instructors to develop a generic rubric (a set of criteria) that could be used to assign partial credit to the distractors in four-part multiple choice questions in a consistent way. In an iterative process the rubric was tested on a series of questions, amended and tested on further questions. The rubric was then used to assign partial credit scores to a General Chemistry examination of 70 items that had been used as the first term test for undergraduates. These partial credit scores were then applied to 1178 scripts that had originally been marked using a dichotomous scoring scheme. The outcomes of the exercise showed that whilst the ranking of high performing and low performing students was almost unchanged, there was some movement in the ranking of students of intermediate performance – students with a partial knowledge of the whole domain being tested were likely to benefit compared with those who have a good knowledge of only part of the domain.

Using such a rubric to ensure the partial credit scores are consistent from item to item would increase the time to develop new test items, but may make for better questions. The authors of the report suggest further work that might be done to determine what further information about students' breadth and depth of understanding can be gained from partial credit scoring.

Ordered multiple choice questions

The extent to which ordered multiple choice (OMC) questions can be used to assess students' level of understanding of core chemistry concepts has been investigated (Hadenfeldt, Bernholt, Liu, Neumann, & Parchmann, 2013). In OMC items the choices represent different levels of understanding of an idea rather than simply 'wrong' answers. The answers students select show where the students are in progressing to a deeper level of understanding.

The researchers used a construct map to describe levels of understanding of the nature of matter, an aspect of students' understanding of chemistry that is well-researched. These levels of understanding were then used to develop 10 OMC items and three free-response items, which were used to check the validity of the OMC instrument. The items were trialled on 294 students in grades 6-12 in a German grammar school.

The scripts were scored using a partial credit model. Item response analysis showed that the OMC items were able to discriminate between three levels of understanding of the structure and composition of matter as effectively as open-ended items. Items such as these have the possibility of providing a powerful tool for formative assessment in the classroom, but could also be used as part of a summative assessment scheme. However, developing these questions builds upon research on students' understanding of the core concepts in chemistry; not all aspects of students' understanding have been researched in such depth as the structure and composition of matter.

Complex multiple choice questions

An alternative partial credit model of scoring uses the 'answer-until-correct' system. In this system students receive feedback on whether their answer is correct, if it is wrong they try again until they select the correct answer; the fewer attempts they make, the higher their score.

Researchers used this model to replace standard multipart short answer physics questions with a series of linked answer-until-correct MCQs, described by the researchers as 'integrated testlets' (Slepkov & Shiell, 2014).

The testlets were developed from short answer questions (SAQ) that formed part of the mid-term and end of term exams in an introductory physics course in a Canadian university. In a traditional SAQ that requires some problem solving students use some answers from earlier parts of the question to answer subsequent parts. In the replacement MCQ testlet the question was broken down into parts, with each part using the answer-until-correct format so that students have the correct answer, which will be needed to answer following parts.

Two parallel papers were constructed for each of the mid-term and end of term exams; each complementary examination had an equivalent number of SAQs and testlets and covered identical

course material, but swapped question formats for each topic covered so that performance on the two approaches to assessing the same idea could be compared. Although there was limited data, as only eight questions were used in this trial, it was found that there was a correlation between the scores on the two types of questions; however the integrated testlets were not as discriminating as the SAQs. Whilst the testlets can be machine-marked, measures of inter-rater reliability showed that there was some variability between scorers on the SAQs.

The testlets provide information about how a student progresses through a problem, but they do not measure exactly the same construct as the equivalent SAQ because the procedural route through the problem is defined by the question, with no room for students to make decisions about how to tackle the task.

Discouraging guessing

When administering MCQs, it is inevitable that candidates may able to score a certain percentage of marks simply by guessing. In a study at the US Naval Academy in Maryland, USA, undergraduate mid-term examination multiple choice papers were amended to decrease the contribution of the final score resulting from guessing (Campbell, 2015). The number of distractors was varied depending on the type of question – those that required extensive reading were limited to four choices, whilst those that gave the answer to calculations had up to 10 possible responses. Students were encouraged not to guess by being told that they would score 4 marks for each correct answer and 1 mark for any answers left bank, thus it was unlikely students would improve their score by uninformed guessing.

It was found that crediting blank answers led to fewer students making uninformed guesses. Reducing guessing increased the reliability of the scores of individual students by reducing the contribution of marks obtained through uninformed guesses to overall scores.

4.1.2 Current use of multiple choice questions to assess chemistry subject knowledge in secondary education in England

All the assessments of chemistry at Advanced GCE (from 2017) and GCSE (from 2018) include some fixed response questions, with most specifications including a section of MCQ within some question papers. These questions usually have a key and three distractors and are scored dichotomously.

4.1.3 Practical considerations in using multiple choice questions

Key informants reported that there was an important place for multiple choice questions in the assessment of chemistry subject knowledge, particularly in contexts where there are large numbers of students undertaking the assessment. MCQs were described as a reliable way of assessing the full examination specification with the facility of quick marking, often by machine, which allows processing thousands of entries in a cost and time effective way.

There were differences reported in terms of how familiar students are with these items. In contexts where MCQs are widely used, it was seen as an advantage that students knew how to handle these questions, but where they were unfamiliar with the format it was noted that students need more examination preparation including guidance about how long they should expect to spend on each question.

In common with the literature cited above, key informants identified a range of practical challenges associated with use of MCQs:

- the possibility that students answer questions correctly by chance or educated guessing
- difficulty in writing questions (and associated difficulty in recruiting appropriately experienced question setters), particularly in finding suitable distractors that are not misleading, too similar to the correct answer, or which require a lot of processing by students in the time available
- external misconceptions that MCQs are inherently easy

- the possibility that a large amount of reading or processing is needed for a single mark.

It was also identified that standard MCQs are unsuitable for criterion-referenced assessment, and for assessing some objectives, for example, whether a student can create a reasoned argument, work through a multi-step calculation, or draw conclusions from data.

The use of on-screen assessments opens the way for using MCQs for adaptive questioning, where the route taken through the questions is determined by students' responses. If this approach is used for summative assessment, there may be a perceived unfairness associated with students answering different questions, and a different number of questions; this development would require more questions to be supplied by suitably experienced question setters. It was estimated by one of the key informants that this approach would require a bank of over 1000 items, so had been dismissed as impractical given organisational resource constraints.

Other forms of MCQs mentioned by key informants include two-tier questions or a single MCQ with students having to explain their choice.

4.2 Closed response questions

Closed response questions are objective questions where students supply the answer, but the answer is constrained by the question format (Black, 1998). There is a wide range of tasks that can be set using this format, some common examples include:

- supplying a single word or short phrase to a specific knowledge question
- supplying words to complete a cloze passage
- answering a short numerical problem
- completing a chemical equation.

Usually there will be a single mark awarded for the answer.

4.2.1 Developments in the use of closed response questions in chemistry

A number of researchers are investigating the use of concept maps to assess students' understanding of the links between ideas within a particular domain in chemistry. This type of assessment is very open and can be challenging to mark (see later in this report) (Lewis, Shaw, & Freeman, 2011). For this reason researchers have tried to design instruments based on the idea of concept mapping but to generate a more closed assessment.

Vachliotis, Salta, Vasiliou, and Tzougraki (2011) developed what they called Systemic Assessment Questions. These questions used concept mapping techniques to show the relationship between a group of organic compounds and the chemical reactions that linked them. Partially completed diagrams were given to students and they were asked to fill in the spaces and also draw arrows to show the direction of the relationships.

These novel questions were given to 11th grade Greek chemistry students in class tests along with some conventional fixed response questions covering the same domain. Teachers had used similar diagrams in teaching the topic, so students were familiar with the approach. These were closed response questions with only one possible correct answer for each space, thus ensuring that the scoring of the questions can be reliable. 65 students answered the first test and 42 the second test; with only 7 out of the 20 questions over two tests being of the novel kind it is difficult to draw strong conclusions about the potential of this model; further work would be needed to establish whether the approach has a wider application.

4.2.2 Current use of closed response questions to assess chemistry subject knowledge in secondary education in England

All the assessments of chemistry for Advanced GCE (from 2017) and GCSE (from 2018) included some closed response questions. Closed response questions are useful for assessing recall of chemistry ideas, they can also be used to assess students ability to apply an idea in a different context from which is may have been taught.

At Advanced GCE closed questions are used to assess such ideas as:

- electron configuration of a named ion
- bond angle of a complex ion
- ionic equation for a reaction
- systematic name of a compound from its formula
- identify a functional group on a diagram of the structural formula
- extracting data from a graph
- identify the type of reaction described.

(AQA, 2015; Edexcel, 2015a; OCR, 2016b)

Closed questions are also used at GCSE where they might ask students, for example, to :

- complete a table that has some data provided
- use data about boiling point and melting point to deduce the state of a substance at a given temperature
- complete a sentence describing a reaction
- complete a dot and cross diagram for a covalent compound
- write a word equation for a reaction
- complete a diagram showing paper chromatography
- calculate a mean value from data.

(AQA, 2016; Edexcel, 2016; OCR, 2016a)

4.2.3 Practical considerations in using closed response questions

Closed answer questions have tightly defined mark schemes; this means scoring can be reliable and gives the possibility of on-screen assessment with computer scoring. Whilst the required answers are tightly defined there may need to be some judgement made about the acceptability of the answer, for example whether the correct spelling is required, and if not how much leeway is allowed in interpreting the given spelling. If the questions are to be scored by computer all possible acceptable answers would need to be supplied.

4.3 Open response questions

Open response questions (also called constructed or free response) are those questions that do not constrain the student's response. The response may be written, drawn, or a calculation, or a combination of these. The score for the questions may be one or two marks or many marks for an essay or other extended piece.

This section reports on the variety of formats for open response questions used in examination contexts, the following section considers free response tasks that are carried out by students during the course, before the examinations, (usually) marked by their teachers, and the mark combined with those from written papers.

4.3.1 Evidence from the literature about formats of open response questions

Short answer questions

In this context short answer questions (SAQ) are those that require a short textual answer, a diagram or calculation, and are usually worth 1-4 marks. SAQ are currently marked by examiners, rather than machine-scoring, which is used for multiple choice questions (MCQ). The use of machine scoring to score open response questions is a developing area of research, discussed *in Section 6: Future directions*. As machine-scoring of MCQs is more reliable than hand-marking SAQ it needs to be considered what advantages SAQ bring to an assessment framework.

Often SAQ are part of a structured multi-part question which requires students to address a more complex problem that has been broken down into a series of steps. These questions allow the

student more freedom of expression than the closed response questions described earlier and consequently can reveal more about the students' thinking.

Hudson and Treagust (2013) investigated an apparent difference in performance by male and female students in the Victoria (Australia) state university entrance examinations for chemistry. The examinations consist of MCQ and SAQ assessing both recall and application of chemistry knowledge to a particular situation, usually requiring a calculation. The researchers developed sets of paired questions that assessed the same content knowledge in MCQ or SAQ format. The tests were given to students in four secondary schools that have traditionally performed well in the university entrance examinations.

The results of the study showed that within this cohort the lower ability students found the SAQ questions slightly more difficult than the MCQ. It has been argued that male students perform better on MCQ questions than female students however this study found that, after student ability is taken into account, there was no significant difference in the performance of male and female students on the two question formats.

In a study to find out what would be the effect of removing SAQ from high stakes assessments Lissitz et al. (2012) used a variety of statistical tools to investigate whether there are differences in the contribution that MCQs and SAQ make to the overall scores in the Maryland High School Assessments (USA). The researchers used the scores of 10,555 students in their high school graduation examinations in biology, English, algebra, and government. The SAQs in the tests are worth 3 or 4 marks. In all four tests MCQs and SAQs were designed to test the same knowledge and skills. In each paper approximately one third of the marks were for SAQs.

The researchers found that whilst the MCQs and SAQs did measure broadly the same constructs, the tests that included SAQs were harder than those just including MCQs. They also found that female students' performance was higher relative to males when the test included SAQs, providing support for the theory that females' high verbal abilities result in higher scores on SAQ items. This result contrasts with the findings of Hudson and Treagust (2013) above, however the cohort in the Australian study did not include the full ability range, and the difference in performance on MCQ and SAQ is more evident amongst lower ability students who may have difficulty in the writing demands of SAQ and are also more likely to benefit from scoring by guessing on MCQ. Another difference between the two studies is that the SAQ were different in style between the two studies, with the Australian study being confined to mostly quantitative problems.

Lissitz et al. (2012) point out that although there is a good correlation between scores on MCQ and SAQ they do not test exactly the same constructs and without SAQ items in the assessment the focus of teaching may narrow to only those skills required to do well in a MCQ examination.

Concept mapping

Concepts maps are used in teaching to help students make connections between ideas; studies have found that they are useful tools to help students generate meaningful connections between chemical concepts (Francisco, Nakhleh, Nurrenbern, & Miller, 2002). It is difficult to develop a consistent model for writing scoring rubrics for students' constructed concept maps because of the variety of ways in which students present their ideas and the connections between them.

Chemistry tutors at Kennesaw State University (USA) developed the idea of 'creative exercises' in which students are given a brief prompt about a chemistry context and asked to write down as many distinct, correct, and relevant, statements as possible (Lewis et al., 2011). The statements expected are of the sort that would be shown on a concept map for the prompt context. The exercises were used as a teaching tool and also incorporated into in-class tests. The tests were taken by 276 students and were graded by three people. The researchers compared the students' scores on the tests with their scores on First-Term General Chemistry examination of the American Chemical Society (ACS) Examination Institute. The exercises and examination both cover the same chemistry content, but the ACS is a multiple choice examination, so it would not be expected that there would

perfect agreement between the two; the correlation between the scores on the exercise done as in-class tests and the ACS examination scores was 0.50. The researchers proposed that the combination of the acceptable correlation and the satisfactory inter-rater agreement, leading to the conclusion that this might be a question format that is generalisable to other situations.

Open book examinations

Whilst many chemistry examinations will allow the use of a data book, open book assessments are those where students may use textbooks, notes, or other reference materials in an examination. There are variations in this type of assessment: for example students are supplied with an unmarked copy of book (such as the set text in an English literature examination); or students may only be able to take their own hand-written notes; or they may be able to annotate their textbook.

The rationale for introducing open book examinations includes reducing examination stress, reducing the need for rote learning, and encouraging higher level thinking skills. Many of the more recent reports in the literature describe single studies by instructors working in a higher education setting, however there has been some work done with teachers in schools.

A reform in the curriculum in Norway in 1994 placed a new emphasis on school-based evaluation and also the promotion of higher level cognitive skills. Researchers took this reform as an opportunity to work with high school science teachers in developing a teaching and learning framework that supported these aims; they included open book assessment in the development (Eilertsen & Valdermo, 2000). Over the three years of the project researchers found that both students and teachers developed a better understanding of the nature of knowledge and the process of learning. The study showed that the majority of students showed improved learning and recognised that, even though the tests were perhaps more demanding than closed book tests, they did lead to better learning.

An open book examination developed at the Department of Mechanical and Mechatronic Engineering at the University of Sydney, Australia, was described as a 'Power Test' (Baillie & Toohey, 1997). This examination took a very open approach – students were able to take up to 8 hours for the examination; they would write their paper in an examination room but were allowed out to visit the library, collaborate with colleagues, and take comfort breaks. They could not take anything in or out of the examination room. The purpose of this approach was to remove the need to rote learning and encourage deep learning. Much thought went into designing assessments that tested this deep learning.

Students were prepared for this different assessment approach through the teaching and a practice examination. The examinations were marked using the SOLO taxonomy (Biggs & Collis, 1982) and marks were compared to the students' marks on a closed book examination taken the previous year. The open book approach showed that more students were able to achieve the higher grades, demonstrating the higher level thinking skills described in the SOLO taxonomy as 'extended abstract'.

Both the studies reported above emphasise the need to prepare students for this different kind of assessment and to develop an assessment appropriate to the approach. This necessity is reflected in a study by Moore and Jensen (2007) in which they gave the same examination to two halves of a cohort of 351 undergraduate students on an introductory biology course at US university. Those students who knew they would have access to their books in two of the mid-term tests (the experimental group) did not score as well on those tests as those who took the identical closed papers (the control group), neither did the experimental group score as well on the final closed book examination that covered all the content of the course. The students had identical preparation for the tests and the same opportunities for additional support; this support was taken up far more by the control group than by the experimental group.

A study at the US Air Force Academy in Colorado, USA, investigated the effect of open-book exams on student achievement in an introductory statistics course (Block, 2012). The course is designed to

improve the students' conceptual understanding of statistics and analysis. Prior to the first use of the open book examination teachers emphasised to students that the examination would be assessing deeper thinking skills and encouraging students to prepare for that. However in the first trial of the open book examination students' scores were lower than the previous cohort who had taken the same examination under closed book conditions; it appeared that they had relied too heavily on being able to find answers in the textbook. For the following cohort instructors spent more time emphasising that the test would be hard and that they needed to prepare for it even though it was open book. Students did before better than the previous cohort but didn't like being told that the test was harder (it was the same test). For the following cohort students could take into the examination their own hand-written notecards, but no text book. This led to better preparation and performance – and happier students.

All of these studies suggest that there is a role for the open book examination in testing higher level thinking skills, leading to deeper learning, providing that the teaching and assessment are aligned in that direction. This finding is supported by earlier studies reported by Cresswell (2000). The material that can be taken into the examination needs to be considered carefully – too much material may lead students to rely on it and waste time hunting for answers. Hand-constructed notes help students prepare for the examination and also have a useful purpose in helping students think about their learning.

4.3.2 Marking open response questions

Examination questions and mark schemes are written by experienced examiners, many of whom are, or have been teachers. The question papers and mark schemes are also used by teachers as a guide for what is expected in future examinations. Mark schemes used in the current GCSE and Advanced GCE chemistry examinations in England are currently either points-based schemes or levels-based schemes.

Points-based mark schemes are currently used for most open response questions. There may be a one-to-one correspondence between marks for the question and points listed in the mark scheme, or the mark scheme may give more alternatives than there are marks.

Levels-based mark schemes are used to mark longer prose answers – from one or two paragraphs up to extended essays. The mark scheme describes a number of levels of response, each with an associated band of marks. The description for each band will identify the criteria a candidate's response needs to match to be in that band. Normally examiners apply a principle of 'best fit' to decide the mark to award. This type of mark scheme is also called a level of response (LOR) mark scheme, or a banded mark scheme.

The awarding bodies and Ofqual have carried out a number of studies to evaluate the factors that affect reliability of marking of different question types. The outcomes from the studies indicate that an examination paper that uses closed questions and questions that can be marked using points-based mark schemes might yield more reliable marking. However there are some aspects of learning which do not lend themselves to these question styles. For high-tariff questions that demand an extended open response, a levels-based mark scheme may be more appropriate. To not include questions of this type would reduce the construct validity of the assessment, that is, the assessment would not be able to assess all the knowledge, understanding, and aptitudes that the qualification is expected to reflect. Thus there is a tension between ensuring the marking is reliable and producing a valid assessment (Ofqual, 2014b).

Possible ways of improving reliability include using computer-based scoring systems, or comparative judgement approaches; these will be discussed in *Section 6: Future directions*.

Criterion-referenced marking

Education reforms regularly result in the requirement to develop the higher level thinking skills of students. For example, the PISA 2015 cognitive framework described high cognitive demand:

> Analyse complex information or data; synthesise or evaluate evidence; justify; reason, given various sources; develop a plan or sequence of steps to approach a problem. (OECD, 2016, p. 39).

There are concerns expressed that most assessment practices focus on lower order thinking skills, which in turn leads to shallow learning. In a study of the assessment instruments used in four Australian states Fensham and Bellocchi (2013) sought to evaluate how well each instrument was aligned to, and facilitated, the higher order thinking skills described in curriculum documents. They found that the state assessments of Queensland, which used a common set of criteria to score all assessment tasks, provided better opportunities for assessment and crediting of higher order thinking skills than assessments that used a points-based marking system.

There have been a number of researchers who have developed theoretically-informed frameworks that identify levels of understanding in particular areas of science. These frameworks could be used to develop criterion-referenced marking rubrics for a variety of tasks within the domain of the framework.

Section 4.1 described the work of Hadenfeldt et al (2013) who used a theoretical framework based on knowledge about progression in students' understanding of particular ideas to develop ordered multiple choice questions; this section reports on work to develop frameworks that have a broader use than a single scientific idea.

Curriculum reform in Germany in 2004 resulted in national educational standards based on competence levels, which defined the learning outcomes for 10th grade students, without describing compulsory content. Walpuski, Ropohl, and Sumfleth (2011) developed a model to describe increasing demands of tasks designed to provide evidence of the area of competence 'application of content knowledge'. These descriptors were tested using multiple choice questions.

Bernholt and Parchmann (2011) describe a hierarchy with five levels of complexity that could be used to develop tasks and subsequently assess students' achievement in the domain of chemistry content knowledge. The hierarchy describes increasing levels of interconnectedness between ideas, from the first level where students base their explanations on everyday life, are able to make observations and give examples of phenomena. At the highest level students are able to explain nonlinear relationships and handle several variables and their contribution to complex cause-effect relationships.

Whitehouse (2014) used research about teaching argumentation (Erduran, Simon, & Osborne, 2004) to develop a theoretical framework that could be used to write levels-based mark schemes for questions that require students to give a scientific explanation for a phenomenon or provide an argument. This framework was specifically developed to match the requirements of the GCSE assessment regime.

All the frameworks described above would be useful in developing the criterion-referenced mark schemes that Fensham and Bellocchi (2013) suggest can lead to assessment of higher level thinking skills.

4.3.3 Current and past use of open response questions to assess science subject knowledge in examinations in secondary education in the UK

All the assessments of chemistry in Advanced GCE (from 2017) and GCSE (from 2018) include open response questions. Many of the questions are structured questions with tariffs from 1 to 4 marks. The maximum mark for a question or part question is currently 6 marks. Some of these 6 mark questions are marked using levels-based marking.

The Scottish Qualifications Authority (SQA) Scottish Highers and Advanced Highers examinations include open response questions described as 'open-ended questions (SQA, 2010). These questions require students to solve a problem or challenge, drawing on their understanding of key chemical principles. The questions have no unique correct answer; students can take several different routes

to a good answer and are rewarded for analysis and creativity. Marking of these items uses a levels-based mark scheme awarding 0-3 marks, allowing students to be rewarded for chemical insight and holistic understanding of the subject. These items are thought to discriminate well. This type of question is challenging for students who are more accustomed to memorisation and recall, and guidance needs to be given to students on the time they should expect to spend on these questions as there is a risk that they could devote a disproportional amount of time in the examination to solving these problems.

A variety of formats of open response questions have been in used examination papers to assess science subject knowledge within GCSE and Advanced GCE examinations in the past, however in recent rounds of curriculum reform the variety has diminished as Ofqual has sought to ensure that it is straightforward to demonstrate comparability of grades between different awarding bodies. Many of the innovative assessment instruments were developed as part of curriculum development projects to ensure that the examinations reflected the main aims of the learning. For example, Black (1998) relates the six components of the original Nuffield Advanced Physics assessment, including four different written papers, to the aims of learning they were attempting to evidence.

Comprehension passages

Questions based around a comprehension passage have been a feature of a number of assessments in science in the past. In the Nuffield A-Level Chemistry examination there was a passage about a modern aspect of chemistry, which students would read in the examination and then answer questions. What was strongly discriminating was the task of making a précis of some aspect of the passage in 100 to 150 words.

A variation on the straightforward comprehension passage has been to provide the passage in advance as pre-release material. This has the advantage of allowing the examiner to set a more substantial amount of text and data as students have the opportunity to read it in advance and research any ideas that they do not understand. A paper including pre-release material was used in assessments of Twenty First Century Science GCSE Chemistry (OCR, 2005) until 2012 and is currently included in Salters A-level Chemistry (OCR, 2014b) and A-level Science in Society (AQA, 2013). In a study of the impact of including pre-release material in a GCSE geography examination Johnson and Crisp (2009) found that this form of assessment had a positive effect on teaching and learning and that the assessment had good construct relevance, enhancing construct validity.

4.3.4 Practical considerations related to open response questions identified by key informants

A range of different types of open response tasks were identified by the key informants. These included producing a précis of a piece of writing on contemporary chemistry, and descriptions, explanations, or predictions of behaviour of chemical systems. Open response questions allow students to demonstrate knowledge and understanding of the specification, and that they can integrate chemical ideas from different parts of the specification. They can also give students the opportunity to demonstrate creativity, for example, by drawing on evidence to create a new argument or answering in a creative way. Open response questions also require that students communicate in a structured, coherent way, drawing on evidence to create an argument.

Challenges with open response questions include addressing issues related to the reliability of the assessment, as marking extended answers inevitably requires a degree of interpretation.

Although open book assessments are not currently being used for the assessment of chemistry subject knowledge at secondary level, key informants discussed the value of these assessments as they do not prioritise recall or rote memorisation, but rather favour application and understanding. There was some discussion as to whether these assessments could favour faster readers, and whether some students would spend a disproportionate amount of time referring to the text rather than preparing their response.

Pre-release materials have been used in a range of assessments identified by key informants. This format was viewed as helpful in terms of allowing students to compensate for reading difficulties by spending more time with the article in advance of the exam. However, if the texts are quite short, there is a possibility that students will second-guess the examination questions.

Section 5: Performance assessments

The previous section described assessment of chemistry subject knowledge through written papers taken under examination conditions. This section describes other instruments that are used to assess students studying chemistry and considers the degree to which they can provide information about students' understanding of chemistry content. These instruments are often described in the literature as 'performance assessments', these are "assessments of activities which can be direct models of the reality to be assessed rather than disconnected fragments or surrogates" (Black, 1998, p. 87). Often these tasks are assessed by the teacher, followed by a moderation process.

5.1 Using practical work for the summative assessment of chemistry knowledge and understanding

5.1.1 Context

In England, practical work is seen as an essential component of school science courses. However, there is longstanding and ongoing debate about the purposes served by practical work and the extent to which it develops scientific knowledge and understanding (for example, Abrahams, Reiss, & Sharpe, 2013; Hodson, 1996).

The literature on practical work is extensive and wide-ranging. A substantial strand within this focuses on the nature and purposes of practical work. Here, it is clear that practical work is seen as a means of developing scientific knowledge and understanding (Bennett, 2003; Millar, 2004). A further, though less substantial, strand in the literature considers ways in which practical work can be assessed.

A comparatively recent review of the assessment of practical work (Abrahams et al., 2013) noted that assessment of practical work is likely to involve a degree of conceptual understanding. The review itself, however, focused on the assessment of practical skills, reflecting the fact that practical work seldom appears to have been used for the assessment of chemistry knowledge and understanding.

5.1.2 Details from the review

Fourteen publications emerged from the searches as potentially including information on the use of practical work to assess subject knowledge. However, on reading the publications, it became clear that the primary focus was on the assessment of practical skills, with little or no mention being made of assessment of subject knowledge.

In one study, Kirton, Al-Ahmad, and Fergus (2014) reported the use in a first year undergraduate chemistry course of a circus of thirteen five-minute laboratory stations 'to develop and reward competency in the laboratory' (p648). Three of the stations are reported as assessing 'chemical terminology', 'fundamental chemical principles', and 'organic compounds and reactions', but insufficient detail is provided in the publication to know how these dimensions were assessed.

The review also looked at the assessment of subject knowledge through the undertaking of practical independent research projects (IRPs). IRPs are open-ended practical projects where students have a degree of control over the focus of the practical work and the way in which the work is undertaken. Where such work is assessed, this often takes the form of an assessment of a written report on the IRP and/or a presentation on the work.

There are comparatively few examples of IRPs being used to assess conceptual understanding for summative purposes. Primarily this is because students or groups of students in the same class undertaking IRPs are likely to be focusing on different areas of chemistry/science. Thus summative assessment of IRPs tends to focus on the assessment of practical skills.

Five studies, all undertaken in the USA, reported linking the undertaking of IRPs to assessment of conceptual knowledge. In four cases, this involved the use of tests or state datasets, either to compare knowledge before and after undertaking an IRP (Charney et al., 2007; Sikes & Schwartz-

Bloom, 2009), or to compare the performance of groups of students who had undertaken IRPs with those who had not (Krajcik & Blumenfeld, 2006; Sahin, 2013). One study assessed conceptual knowledge directly as part of the IRP (Burgin, Sadler, & Koroly, 2012). In order to take account of the diversity of the IRPs undertaken by students, students were asked to develop concept maps to represent the science understandings related to their research. This was done at the start and end of the IRP to assess they ways in which conceptual understanding had evolved.

5.1.3 Commentary

The review findings suggest that comparatively little use is made of practical work to assess conceptual understanding in chemistry/science. It is clear that the successful undertaking of practical work normally requires appropriate levels of relevant conceptual understanding. However, the challenges posed by undertaking reliable and valid assessment of practical skills appear to restrict the use of practical work to assess conceptual understanding.

5.1.4 Information from key informants about the assessment of practical work

Three different mechanisms for assessing practical work were identified by key informants: coursework, practical examination, and direct teacher assessment of practical skills.

Prior to the reforms to GCSE in England that will lead to new assessments in 2018, assessment of practical work has been carried out through the use of controlled assessments, the format of which was tightly controlled by the awarding bodies. It was suggested by a key informant that this led to the assessment not discriminating well as students were prepared extensively. This observation is supported by data from an Ofqual consultation (2014a).

Practical skills assessment for Advanced GCE has recently been replaced by a 'practical endorsement'. The aim of the assessment is to know that the student is competent to be in a lab, to handle equipment, and conduct a procedure safely. This non-examination component of the qualification is teacher assessed. The approach presents a new set of challenges in terms of ensuring that there is consistency in standards across schools, and in monitoring implementation through observations, interviews, student work, and scrutiny of records. Awarding bodies have such mechanisms in place, including monitoring, moderation and verification to standardise assessments across centres. Comments from key informants included recognition of their validity; there was also a perceived positive impact of the assessment on teaching, with one key informant noting that this should allow teachers to use a wider range of practical work in their teaching.

Some jurisdictions have moved away from a set of *required* practicals towards *suggested* practicals in order to give teachers freedom to make decisions appropriate to their own context. Reasons for this were:

- that a trend had been noted towards memorisation of the required practicals and how to handle questions about these in an examination situation
- at times the impact of the required practical was anti-educational, with teachers encouraged to use the required practical, even when they knew of a better practical to illustrate the intended learning point.

The suggested practicals are linked to the specification chemistry content; examination questions do not assess them directly but require candidates to make decisions, for example, selecting appropriate apparatus or planning procedures.

Practical examinations are used in some jurisdictions, but were considered by informants to be restrictive, given the time constraints, and the need to be able to mark reliably. These factors tended to favour a narrow range of practical tasks being used in the assessments.

Concerns were expressed by a number of key informants relating to teacher assessment of practical work. These concerns were associated with perceptions of trustworthiness, particularly in relation to large, diverse and international cohorts, and also when situated within high stakes assessment and accountability regimes.

5.2 Other performance assessments of chemistry knowledge and understanding

This section reports on two alternative modes of performance assessment that assess conceptual understanding of chemistry identified in the review.

5.2.1 Oral examinations

Dicks et al. (2012) report on the introduction of oral examinations to replace traditional examinations used to assess aspects of organic chemistry in second and third-year undergraduate university courses. Students were required to select a named chemical reaction from a database and then prepare for a 15-minute discussion about the reaction with a panel comprising two course instructors. Students' views of the oral examination were gathered through a Likert scale and free response questionnaire. The results are not reported in detail, but the authors indicated students reported a high level of learning associated with the oral examination. The authors also judged the oral examination to be successful, it promoted promote student creativity, and did not increase demands on staff time.

5.2.2 Web-based video responses

Tierney et al. (2014) report on the use of web-based video as an assessment tool for first and second year undergraduate students' performance in organic chemistry. Students were given three days to produce a video relating to organic chemistry and involving the use of molecular models. The videos had to include questions for other students that could be answered by 'clicker' response. The use of video capture enabled students' responses to be recorded automatically. The videos were assessed by the staff on the course, who looked for students' higher order thinking skills.

The authors report that students enjoyed making the videos and felt it improved their chemical knowledge. The method also permitted rapid electronic feedback to students on concepts they had not fully grasped. They concluded that, although video assessment does take more tutor time, it is a viable additional tool for assessment of chemical knowledge, whilst also enhancing student-tutor interactions.

5.3 Current and past use of performance assessments in chemistry in secondary education in the UK

Performance assessments (commonly described as 'coursework' in the UK) are included in specifications to assess those qualities of a student that do not lend themselves to assessment in a written examination, for example:

- practical skills, such as making measurements and setting up apparatus
- planning and carrying out an investigation
- carrying out research and evaluating the sources of information

Whilst, as mentioned earlier, these tasks require chemistry knowledge when carried out in the context of chemistry, the weighting in the assessment is loaded towards those abilities that cannot be easily assessed in written examinations. Many example of performance assessment instruments allow a significant degree of autonomy for the student in the choice of topic to be studied, thus the chemistry subject knowledge assessed will be different for each student.

Extended practical investigations have been used in the assessment of chemistry at GCSE and Advanced GCE for many years. Until 2007 all GCSEs in the sciences used a common format of assessment of practical investigation, with a weighting of 25%. Whilst students would need to use their knowledge and understanding of chemistry to achieve the highest marks the emphasis was on assessing the process skills of planning and carrying out a practical investigation. Normally all students would carry out an investigation on the same topic, chosen by the teacher. The work was marked by the teacher, ands, after internal moderation within the school, a sample would be sent for moderation by the awarding body.

The Advanced GCE, Salters Advanced Chemistry, included a practical investigation that gave the students significant autonomy in their choice of topic until the introduction of practical endorsement in 2017 (OCR, 2008). Students were assessed in five areas: research, planning and implementation, manipulation, observation and measurement, and conclusions. Whilst there was some reward for the use of chemical knowledge and understanding in the planning and analysis of results, this only contributed a small proportion of the marks.

There have been a variety of extended research tasks in assessment of chemistry. GCSE Chemistry Twenty First Century Science (OCR, 2005) included the opportunity for a case study related to an aspect of chemistry that involves an element of controversy; one strand of the marking criteria required students to use their understanding of chemistry.

At Advanced GCE the Salters Advanced Chemistry specification (OCR, 2000) included a research task stimulated by a set of articles on a topic of current chemical interest. Students were required to research the topic of the article and write a report outside class-time over a two week period; the essay was marked by an external examiner.

In Scotland, all students of Higher Chemistry or Advanced Higher Chemistry must complete an independent research project, 'the assignment' (SQA, 2016). One of the criteria for the assessment of the assignment is subject knowledge, with 3 out of 20 marks available for that purpose. This has been part of the assessment of chemistry at Higher level for a long time, and is valued by teachers and universities (UCAS recently commented on the assignment as a strength of the Higher qualification).

5.4 Practical considerations in assessment of performance through coursework

Some types of coursework, such as individual investigations, were seen by key informants as valid assessments that allow students to demonstrate that they can think and act as a scientist, and that they possess a range of skills that are relevant to being a skilled chemist. It was observed that source-based case studies that require students to respond to questions on real-world contexts allow students to demonstrate that they can apply and evaluate information and use sound scientific reasoning.

The main practical challenges associated with coursework (including individual investigations) identified by key informants were the:

- public perception that it is open to interference
- role conflict for teachers as assessors
- high marking and moderation workload for teachers
- high demand of work and time from technicians
- impact on school timetable and space
- need to convince the regulator about certain approaches.

For individual investigations and research tasks, the need to support students doing a range of different topics was an additional practical challenge. In contexts where this was assessed by *viva voce* as well as by report, there was a significant demand on staff time as two experienced examiners were required. Although there are challenges in terms of the reliability of this assessment, a number of mechanisms have been introduced to address this including use of clearly defined rubrics, external moderation, and using more than one examiner with expertise in the field. The role of a culture that supports this type of assessment was considered important by key informants; these types of investigations were reported to be valued, formally or informally, by universities.

Section 6: Future directions

Through the review of the literature and the interviews with key informants, a number of emerging developments in assessment were identified and are summarised below as an indication of possible future directions in assessment practice and research.

6.1 Capitalising on the opportunities of electronic assessment

Drawing upon the various sources of information that have been explored, it can be reasoned that developments over coming years will include capitalising on the opportunities offered by electronic assessment. Key informants noted the versatility and scalability of electronic assessments, which when coupled with improvements in infrastructure and security, offer the potential for more widespread use in the assessment of chemistry.

Electronic approaches have been extensively applied to the development of curriculum materials and formative assessments; for example, the resources developed by the Royal Society of Chemistry and available online. There is, however, scope to develop further the validity and reliability of the assessments embedded within such materials.

The possibilities of electronic assessment are many and varied and are already being implemented; in 2015 the PISA assessments for all subjects were delivered by computer (OECD, 2016). These approaches offer the opportunity to use varied stimuli as the basis for assessment items. This could include the increased use of simulations, visualisations, virtual laboratory contexts, game-based approaches, and virtual and augmented reality. A further opportunity is the digital accreditation of the skills mastered via electronic 'badges', for example in the context of laboratory skills (Seery et al., 2017).

A current barrier to adopting electronic approaches in the UK is access to the necessary information technology systems in terms of both infrastructure and security.

Two examples of electronically-enabled approaches to the marking of assessment tasks are provided below as illustrations of possible directions for developments in practice and research: computer-assisted scoring and comparative judgement (including adaptive comparative judgment).

6.1.1 Computer-assisted scoring

Developments in technology have resulted in a growing area of work on the potential use of computers to score open response items. Much of this work originates in the USA, where multiple-choice questions have predominated in undergraduate courses as a means of assessing scientific understanding, particularly where student groups are large. One of the principal motivations for such developments is to provide a means of assessing students' abilities to use evidence and construct arguments – abilities not easily assessed by multiple-choice questions.

Computer-assisted scoring involves experts first coding responses to items. Typically, two experts – 'raters' – code a sample of student responses. These are then checked for agreement, with a third rater assisting where agreement cannot be reached. Once human agreement has reached a high level (e.g. 90%), the remaining responses are then divided between the experts for coding. A substantial portion of the responses generated (e.g. two-thirds) is then used to 'train' a computer marking programme to score students' work. The reliability of the computer marking is then checked against the remaining responses. An increasing number of computer-assisted scoring programmes are becoming available.

Computer-assisted scoring is likely to be a growth area in assessment. Further work is needed to clarify the factors that contribute to high agreement between computer and human marking and the design of items that lend themselves to automated scoring. Its use as a formative assessment tool in teaching through the provision of instant feedback to students and teachers is also an area that is likely to see further work, with this being linked to the provision of on-line support for students.

A potential drawback is that there might be an unwillingness to change a test after deploying the high costs of training the computer to respond to a particular set of questions, although retaining the same test year-on-year allows for comparability between cohorts. The two studies below illustrate how research in the field is developing.

Ha et al. (2011) investigated the use of the Summarization Integrated Development Environment (SIDE) programme for assessing undergraduate students' understanding of evolution using constructed response items. The computer marking model looked for the presence or absence of five key concepts in evolution in students' responses. Ha et al. explored how scoring models built at one university performed at other institutions, how many responses needed to be scored by experts to build scoring models that function effectively across institutions, and the factors that limited scoring efficacy and how these could be limited.

Ha et al. concluded that the computer marking model scored the students' constructed responses as accurately and reliably as human markers. They suggest that that the computer marking model works best where student ideas on a particular topic were well-established. Their results indicated that the size of the sample marked by the experts influences the accuracy of marking to some extent, though this varied by concept. They also found that the diversity of linguistic expressions associated with concepts was highly variable, and not easy to predict in advance, and that this could result in differences between human and computer marking.

Liu et al. (2016) used a computer marking system, c-rater-LM, to score eight science explanation items with constructed responses from students. They investigated the accuracy of the system compared with human marking, possible performance differences for subgroups (gender, first language, and routine use of a computer to do homework), and the factors influencing large scoring discrepancies. They established that good levels of agreement between human and computer scores can be obtained. In general, there were major differences among subgroups. There were no gender differences at all, responding in a first language conferred a small advantage in one item, and there were small advantages in two items where students regularly used a computer for homework. The report indicated that around 500 – 1000 responses were needed to build a reliable scoring model. Misspellings and linguistic diversity accounted for the largest discrepancies in human and computer marking as the range of responses programmed into the computer marking model did not always take account of different words or grammatical structures to express similar ideas. They concluded that computer-assisted scoring has value as an assessment tool and offers a viable alternative to the use of multiple-choice questions. It also provides instant feedback to students, whereas there is a time delay associated with human marking. Additionally, it provides a means of comparing students' performance across different teachers.

Much of the research in the area of computer-scoring is taking place in the USA where the large cohorts make hand-marking of open response questions a difficult task; the more markers that are used, the more challenging it is to expect them to all mark to the same standard.

6.1.2 Comparative judgement and adaptive comparative judgement

Comparative judgement (CJ) and adaptive comparative judgement (ACJ) are highly topical in the field of assessment techniques. In educational contexts, ACJ offers the attraction of replacing 'marking' with teachers' professional judgement. It is based on the premise that that someone reading two assignments finds it easier to decide which is the 'better' of the two than to allocate specific marks with reference to specific criteria. The binary outcomes of many such judgements enable the creation of a rank order of scripts from 'best' to 'worst' (or vice versa).

CJ has its origins in Thurstone's method of comparative judgement (Thurstone, 1959). The advent of modern computers has made the application of the underlying statistical principles very easy, greatly increasing its potential as an assessment method in education. Web-based systems allow assessors to view work on-line and automatically record 'scores' (relative judgements). Such a system enables the scores to be re-estimated after each round of judgement. In each round, an assignment is

compared to another with a current similar score, further refining the scale and placing the assignment at a point on the scale.

Research into the use of ACJ as an assessment method indicates it can reliably show the relative quality of each piece of work. In the words of Pollitt (2012), "The judges are asked only to make a *valid* decision about relative quality, yet ACJ achieves extremely high levels of *reliability* …" (p281).

Pollitt (2012) argues that ACJ has a number of advantages. These include little need for training, other than in the use of the web-based system, and providing a better means of reliably and validly assessing assignments that are problematic, such as "long essays in Politics or History where the complex mix of criteria for content and for quality make agreement on marks difficult to reach, and a single marker – or even two – cannot be considered reliable enough for high-stakes assessment." (p293).

CJ and ACJ have the potential to initiate a major change in assessment methods. There is already a body of literature documenting their reliability and validity, and it seems likely that more will be added to this as their use is explored further. The use of CJ and ACJ in peer assessment offers an interesting way of engaging students in critical reflection on their work, although there are concerns about the acceptability of using peer assessment of work in high stakes assessment.

There is a growing literature on CJ and ACJ, much of which has been contributed by those setting examinations, and which focuses on exploring the appropriateness and potential limitations in relation to reliability (see, for example,Bramley, 2015). Two studies illustrating the use of CJ and ACJ by practitioners in educational settings are summarised below.

Jones and Alcock (2014) report a study they undertook with first year mathematics undergraduates to explore the use of CJ as a means of peer assessment of work. Students were given a short test of conceptual understanding of multivariable calculus. Then, rather than mark each other's work against assessment criteria, students were asked to judge pairs of scripts against one another without any reference to assessment criteria. No training was provided, other than in the use of the software used to record students' judgements. Inter-rater reliability was investigated by randomly assigning the students to two groups and correlating the two groups' assessments. Validity was investigated by correlating the peers' assessments with expert assessors (maths lecturers and postgraduates), novice assessors (social science postgraduates), and marks from other course tests. Interviews were also conducted with some members of each group of assessors to explore their thought processes.

The researchers report high levels of validity and inter-rater reliability in all groups, and suggest that students perform well as peer assessors. They also conclude that use of CJ in this way offers students an opportunity to reflect on their conceptual understanding and ability to communicate mathematical ideas.

In a related study, CJ was used as a means of peer assessment with students aged 13-15 in mathematics classes in three high schools (Jones & Wheadon, 2015). As with the work with undergraduate students, they report that the use of CJ enables high school students to perform well as peer assessors. They suggest that CJ offers the possibility of summative peer assessment in a range of contexts.

6.2 Fieldwork

The use of fieldwork is common in curriculum areas such as geography, biology and the environmental sciences. The miniaturisation of analytical instrumentation increases the feasibility of including more fieldwork in chemistry, for learning and assessment purposes (Stoodley, Nunez, & Bartz, 2014), and this was identified as a likely future direction in chemistry assessment by one of our key informants.

6.3 Regulation as a factor in assessment development

Although the review did not seek the key informants' views on the regulation of assessment, an overarching theme arose from interviews with some of the UK-based informants related to government policy and the regulation of qualifications.

One of the issues identified was the short period of time that specifications were in use before being subject to change, limiting the extent to which innovation could happen as institutional resource was focused on immediate demands arising from changes in curriculum and/or assessment policy.

A second issue was the need to meet the requirements of the regulator in England (Ofqual), and the limited extent to which awarding bodies felt able to make their own decisions. Key informants discussed criteria that had to be met in terms of how assessments are conducted, down to criteria for extended response questions, coverage of assessment objectives, and the proportion of marks that can be awarded for knowledge and understanding, application and interpretation. This was described as being highly restrictive. Concern was expressed that the quality of questions was potentially being reduced to meet the demands of the regulator.

The final issue related to the level of justification required to introduce non-standard approaches to assessment, particularly those carried out under non- examination conditions.

These issues combined raise questions about the extent to which innovation is possible in a heavily regulated system.

Section 7: Conclusions and further work

7.1 Conclusions

The sections above describe the methods and findings of a review of the assessment of chemistry subject knowledge for summative purposes. The review involved a survey of the literature from 2006 to 2016, and interviews with key informants who were approached based upon their experience of assessment. Here the findings are summarised with respect to the questions that guided the review.

What are the main summative assessment models for the assessment of chemistry subject knowledge in secondary education?

The assessment tools available to examiners setting written papers fall into three categories: fixed response questions, where students select an answer from those supplied by the examiner; closed response questions where there is limited scope for student choice – the answer is well defined; and open response questions where there is scope for extended writing.

Open response questions provide scope for the examiner to set a variety of tasks, where students are required to use higher level thinking skills including analysing complex information, synthesising information and justifying conclusions. Many open response questions assessing these higher level skills will include stimulus material, such as text, data or a graph; this information may be supplied in advance of the examination so that students can study the information in advance of seeing the questions. If written appropriately, fixed and closed response items can also test some higher level thinking skills.

Most assessment systems will use a combination of different question types, using the strengths of each to ensure the reported outcomes of the whole assessment provide a picture of students' abilities that is as representative as possible.

Although some performance assessments such as practical work and independent research projects are widely used in chemistry, there is little evidence in the literature that their primary purpose is to assess subject knowledge. Oral assessments, in contrast, were found to be used to assess subject knowledge, although these assessments are less practical to implement for large cohorts.

What are the advantages and disadvantages of each model that are reported in the literature?

Fixed response questions and closed response questions can be marked reliably, particularly when machine marking is used; machine marking is also cost-effective and results can be published quickly. Traditionally such questions are regarded as 'easy', although this is not the case, with multiple choice questions being used at all educational levels to test a range of thinking skills.

Developing high quality fixed response questions, and in particular multiple choice questions (MCQ) is challenging, and ideally questions would be pretested before being used in high stakes assessments. In some jurisdictions the questions are not released to the public and the same test can be used year-on-year, allowing for comparison of cohorts, however this removes the formative opportunities that past papers afford to teachers.

Fixed response and closed response questions are suitable for assessing recall and the ability of students to apply ideas in new contexts; however they do not lend themselves to assessing all higher level thinking skills.

Open response questions provide scope for testing students' abilities to apply their chemical understanding in a range of contexts, using higher level thinking skills. Open response questions can be set at all levels of demand, and either marked using a points-based marking system or using a criterion-referenced system of scoring. A criterion-referenced system allows differentiation by outcome; the same task can be taken by students with a range of abilities, and the criteria are used

to assign a score at the end of the task. By publishing the criteria to teachers and students the requirements of the assessment become transparent; this supports teaching and learning.

A concern has been expressed about the reliability of the marking of open response questions; it has been reported that the more 'open' the question and the more marks that are available the weaker the inter-rater reliability. This concern about reliability may be a limiting factor in the number of marks made available for a single question. Comparative judgement using a criterion-referenced scheme and computer-scoring are two current developments aimed at improving the reliability of marking systems, although these too face challenges in terms of how they will be perceived by students, parents and teachers.

Practical work and extended projects were not found to be routinely used to assess chemistry subject knowledge, but rather a different set of skills. These typically demanded more from schools in terms of space, timetabling, and teacher and technician time, and presented challenges for marking, moderation and standardisation. These challenges are less in some jurisdictions where these approaches are well established as part of the culture of science teaching and assessment.

Oral assessments have been used to assess subject knowledge and were reported to support learning and promote creativity. To ensure that these are valid and reliable, some systems required more than one examiner, the use of rubrics to guide the marking, and the use of internal and external moderation.

What is the depth of research understanding available about the effectiveness of each assessment model? (e.g. is the evidence about each assessment model robust or scarce?)

The scope of the literature review was limited to publications published between 2006 and 2016 that considered summative assessment of the chemistry knowledge of students aged 14-19. A small number of publications outside that range are included in the review to cover aspects of assessment for which there has been no recent work.

There has been limited large-scale research into summative assessment of chemistry specifically. There has been significant research into the psychometric properties of MCQs, and the reliability and validity of such questions. Much of this understanding can be transferred to the development of other forms of assessment.

The review has found that research involving the assessment of chemistry subject knowledge could be considered to fall into four broad categories:

1. Research concerned with developing diagnostic assessments to develop the understanding of students' conceptions of chemical ideas; this research is not concerned with developing assessments to be used for summative purposes and so does not consider their effectiveness in this way – many of these diagnostic assessments involve MCQ.
2. Research, including that carried out by the awarding bodies, which focuses upon a particular model of assessment, and it may be coincidental that the subject being assessed is chemistry. This research is often about reliability of systems of assessment rather than about the validity of the assessment.
3. Research aimed at developing summative assessment tools that can effectively evidence students' understanding of chemistry, for example work aimed at developing MCQs that assess learning beyond factual recall, or the development of new forms of assessment.
4. Small-scale action research projects undertaken in a small number of institutions (often a single institution) and aimed at assessing the effectiveness of a particular form of assessment such as the use of oral examinations in chemistry; this research generally takes place in undergraduate programmes.

The majority of the publications reviewed in this report were published in peer-reviewed journals; many are characteristic of the fourth category above.

The depth and quality of the research evidence base is variable. Research into the refinement of particular approaches (e.g. MCQs, computer assisted scoring) appears to be generating a solid

evidence base. Research into the use of new assessment methods (e.g. comparative judgement, oral examinations) is often characterised by work being undertaken by strong advocates for a particular method. Where the work involves only one (or two institutions) the evidence base is slight.

7.2 Further work

Through the review of the literature and the interviews with key informants, a number of areas of interest falling outside of the scope of this work were identified. These could form the basis of further work in this area and are summarised below.

7.2.1 Disciplinary dimensions

The interviews with key informants touched upon the particular challenges associated with assessing chemistry, as well as the affordances offered by the discipline; for example, two key informants reported that chemistry lends itself to synopticity. Challenges reported included:

A need to supporting learners to competently invoke and switch between the invisible/microscopic domain, the visible/macroscopic domain, and symbolic representations during assessment tasks;

- a need to consider the use of models in assessment tasks: It is necessary to determine whether content knowledge itself is being tested, or the understanding of a model describing the content in question
- the particular demands that the use of chemical symbolism places on electronic assessment and the associated software
- the necessity for learners of mastering certain mathematical skills in order to be able to demonstrate their understanding of a chemical process
- the existence of anomalies in terms of chemical phenomena; there is a need to carefully examine assessment questions for inclusion of unexpected 'exceptions to rules'.

Further work could explore these disciplinary dimensions in greater depth, and compare and contrast these with assessment in other scientific disciplines.

7.2.2 International dimensions

A separate study could explore international systems of assessment with a focus on the assessment of chemistry subject knowledge. For example, the development and adoption of the Next Generation Science Standards (NGSS) in the USA in recent years has led to work relating to the effective assessment in science for learners of school age; the Stanford NGSS Assessment Project Team (SNAP) has released several reports likely to be informative (see for example, Wertheim et al., 2016).

Development work and research centred on international assessments such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMMS) represents a further body of work that could form the basis for an internationally-focused review (see for example, Bybee, McCrae, & Laurie, 2009; Fensham & Bellocchi, 2013; Martin, Mullis, & Hooper, 2016; OECD, 2016).

7.2.3 Other areas of potential interest

Vocational qualifications

While this study has not focussed upon assessment relating to vocational qualifications, this represents an area likely to be of interest as part of further work. For example, the Oxford Cambridge and RSA awarding body (OCR) offers a range of vocational qualifications, including the recently launched Cambridge Nationals and Cambridge Technicals, which include science-related content. Evidence about the effects of assessment practices within vocational qualifications is worth exploring.

'Extension' assessments

A number of other Level 3 qualifications include a requirement for a research project however the topic for the research is of the students choosing so is not necessarily chemistry; these include the Pre-U Independent Research Report (CIE, 2016), the International Baccalaureate Extended Essay (IB, 2017), and the Extended Project Qualification (AQA, 2017; Edexcel, 2015b; OCR, 2014a).

Programmes such as the Royal Society of Chemistry UK Chemistry Olympiad and the Cambridge Chemistry Challenge have a particular focus in terms of assessing chemistry knowledge and skills. These assessments could be considered to be more directly focussed on creative problem solving than is the case for other national examinations.

An exploration of the assessment practices and research evidence relating to these programmes may yield insights of relevance to national examinations.

Experiences of higher education institutions

Owing to the nature of the institution and the student body that they serve, the Open University, UK, is likely to have extensive experience of online and on-screen assessment. Further work focused on these assessment approaches would benefit from an exploration of approaches and taken by this institution.

Approaches and evidence relating to assessments used for admissions to certain university courses may also be of interest, for example the Natural Sciences Admissions Assessment offered by the University of Cambridge. This assessment is designed to 'distinguish across [the] field of high-calibre applicants' to some university programmes (University of Cambridge, 2017). The 'Thinking Skills Assessment' is required to study Chemistry at the University of Oxford, UK, as well as several other degree programmes, and aims to assess discipline-independent problem-solving skills and critical thinking skills (Cambridge Assessment, 2017).

Higher education institutions may also offer a useful perspective in terms of their expectations and experience of the readiness of their incoming students to engage with learning and assessment at the tertiary level.

Appendix 1: Search strategy

Focus

Articles, reports and other publications on the summative assessment of chemistry subject knowledge in secondary education

Population

School and college students aged 14-19

Limits

Published in English between 2006 and 2016

Results of searches of ERIC databases

Search period: 11th January to 20th January 2017

Total records retrieved: 1,148

Search terms

chemistry AND assessment AND high school NOT formative assessment (114 records)

chemistry AND test AND high school NOT formative assessment (268 records)

chemistry AND multiple choice (68 records)

chemistry AND standardised tests (33 records)

chemistry AND summative assessment (20 records)

chemistry AND teacher assessment (31 records)

chemistry AND assessment tools (46 records)

chemistry AND evaluation methods (146 records)

chemistry AND assessment AND science teaching (53 records)

chemistry AND test AND science teacher (75 records)

chemistry AND practical assessment (4 records)

chemistry AND examination AND school (102 records)

chemistry AND open-book exams (0 records)

science AND open-book exams (1 record)

chemistry AND short answer questions (42 records)

science AND assessment tools (143 records)

comparative judgement AND summative assessment (2 records)

Notes

Use of thesaurus facility in search allowed meant search for 'school' included 'college'.

The Social Science Citation Index and PsychINFO yielded records that duplicated those found in ERIC and BEI.

Additional publications identified though hand-searches and recommendations from key informants were added to those identified through the electronic searches.

All the publications identified through electronic searches were imported into an EndNote database, and from there to a Word document. At that point details of the additional publications were added.

Appendix 2: Inclusion criteria

Publications will be included in the review, subject to the exclusion criteria below, if they address one or more of the review research questions:

1. What are the main summative assessment models for the assessment of chemistry subject knowledge in secondary education?
2. What are the advantages and disadvantages of each model that are reported in the literature?
3. What is the depth of research understanding available about the effectiveness of each assessment model (e.g. is the evidence about each assessment model robust or scarce?)

Publications will be excluded from the review on the basis of the following criteria:

1. Published before 2006
2. Do not focus on students aged 14-19
3. Do not focus on chemistry
4. Do not cover aspects of summative assessment of chemistry subject knowledge

A degree of professional judgement was required in applying the criteria; where there were publications reporting research into the use of a summative assessment instrument used in a discipline other than chemistry, these were included in the review if it was thought that the same approach might be applied to assessing chemistry subject knowledge. Similarly, although the first research question refers to secondary education, some interesting uses of assessment instruments in undergraduate chemistry courses are included.

Appendix 3: Data extraction form

Author(s)	
Year	
Title	
Source of publication[1]	
Key words:	
Country	
Assessment type	
Subject assessed	
Age of learners	
Type of work	
Abstract	
Comments	
Details of researchers[2]	
Name of assessment programme (if applicable)[3]	
Brief description of assessment	
Relevance to chemistry – what kind of knowledge being assessed?	
Advantages / disadvantages	
Anything else worth noting?	
Where publication includes a research study	
Aims of study	
Summary of study design, including details of sample	
Methods used to collect data	
Data collection instruments, including details of checks on reliability and validity	
Methods used to analyse data, including details of checks on reliability and validity	
Summary of results	
Conclusions	
Links with any other publications in review?[4]	
Anything else worth noting?	

[1] e.g. name of journal, weblink if online
[2] In particular, what is the relationship of the report authors to the work being reported?
[3] e.g. GCE Advanced level
[4] e.g. publications by same author, or research on same assessment approach

Appendix 4: Key informants interview schedule

Interviews are intended to probe:

- Current or past uses of various assessment types to assess chemistry subject knowledge
- Information on assessment types not currently used in chemistry, for ages 15-19
- Practical advantages and disadvantages of various assessment types
- New/unpublished research that we might not be aware of

We are particularly interested in:

- Ages 15 to 19
- Assessment of subject knowledge
- Assessment used summatively

The Royal Society of Chemistry (RSC) is interested in various types of assessments used summatively with respect to chemistry subject knowledge at upper high-school level (ages 15-19).

- **Please would you tell us briefly about your role/experience of working in the area of assessment?**

We are interested in current or past uses of various assessment types to assess chemistry subject knowledge.

- **Could you please provide an overview of some of the main types of assessments currently used by your organization / in your context/country to assess chemistry subject knowledge for summative purposes?**
- **Do you have any experience of the types of assessment on this page (a list of assessment types with brief descriptions was provided to informants in advance of interviews). Would you suggest any others for us to consider?**

We are interested in assessment types not currently used in chemistry, for ages 15-19.

- **Are you aware of any particular approaches used in other subjects or for other age groups, but not currently applied within chemistry?**
- **Are you aware of any non-traditional or emerging approaches that might be relevant to the assessment of subject knowledge?**

We are interested in the practical advantages and disadvantages of particular forms of summative assessment in chemistry.

- **What would you see as the practical advantages or disadvantages of any particular approaches to assessment?**
    - *For example, are there approaches likely to be effective for assessing knowledge, but which are not practical to implement?*
    - *Are there approaches that are adopted for practical reasons for which concerns over validity or reliability exist?*
    - *Would you have any comments from the point of view of learners?*
    - *Would you have any comments from the point of view of examiners or schools?*
    - *Would you have any comments from the points of teachers?*

Part of our study involves looking at what has been written or researched about different types of assessment.

- Is there anything you would particularly recommend we should look at? For example very recent or unpublished research, or data collected for internal use?

We also want to make sure we talk to key people such as you about assessment of subject knowledge.

- Who do you think it is essential we consult?

Is there anything else you would like to add?

Thank you for your time.

Appendix 5: Publications included in the review

Baillie, C., & Toohey, S. (1997). The 'Power Test': its impact on student learning in a materials science course for engineering students. *Assessment & Evaluation in Higher Education, 22*(1), 33-48.

Bernholt, S., & Parchmann, I. (2011). Assessing the complexity of students' knowledge in chemistry. *Chemistry Education Research and Practice, 12*(2), 167-173.

Block, R. M. (2012). A Discussion of the Effect of Open-book and Closed-book Exams on Student Achievement in an Introductory Statistics Course. *PRIMUS, 22*(3), 228-238.

Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgement*. Retrieved from Cambridge, UK: http://www.cambridgeassessmentjobs.org/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf

Burgin, S. R., Sadler, T. D., & Koroly, M. J. (2012). High school student participation in scientific research apprenticeships: Variation in and relationships among student experiences and outcomes. *Research in Science Education, 42*(3), 439-467.

Campbell, M. L. (2015). Multiple-Choice Exams and Guessing: Results from a One-Year Study of General Chemistry Tests Designed to Discourage Guessing. *Journal of Chemical Education, 92*(7), 1194-1200.

Charney, J., Hmelo-Silver, C. E., Sofer, W., Neigeborn, L., Coletta, S., & Nemeroff, M. (2007). Cognitive apprenticeship in science through immersion in laboratory practices. *International Journal of Science Education, 29*(2), 195-213.

Dicks, A. P., Lautens, M., Koroluk, K. J., & Skonieczny, S. (2012). Undergraduate Oral Examinations in a University Organic Chemistry Curriculum. *Journal of Chemical Education, 89*(12), 1506-1510.

Domyancich, J. M. (2014). The Development of Multiple-Choice Items Consistent with the AP Chemistry Curriculum Framework to More Accurately Assess Deeper Understanding. *Journal of Chemical Education, 91*(9), 1347-1351.

Eilertsen, T. V., & Valdermo, O. (2000). Open-book assessment: A contribution to improved learning? *Studies in Educational Evaluation, 26*(2), 91-103.

Fensham, P. J., & Bellocchi, A. (2013). Higher order thinking in chemistry curriculum and its assessment. *Thinking Skills and Creativity, 10*, 250-264.

Francisco, J. S., Nakhleh, M. B., Nurrenbern, S. C., & Miller, M. L. (2002). Assessing Student Understanding of General Chemistry with Concept Mapping. *Journal of Chemical Education, 79*(2), 248.

Grunert, M. L., Raker, J. R., Murphy, K. L., & Holme, T. A. (2013). Polytomous versus Dichotomous Scoring on Multiple-Choice Examinations: Development of a Rubric for Rating Partial Credit. *Journal of Chemical Education, 90*(10), 1310-1315.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying Computerized-Scoring Models of Written Biological Explanations across Courses and Colleges: Prospects and Limitations. *CBE - Life Sciences Education, 10*(4), 379-393.

Hadenfeldt, J. C., Bernholt, S., Liu, X., Neumann, K., & Parchmann, I. (2013). Using Ordered Multiple-Choice Items to Assess Students' Understanding of the Structure and Composition of Matter. *Journal of Chemical Education, 90*(12), 1602-1608.

Hudson, R. D., & Treagust, D. F. (2013). Which Form of Assessment Provides the Best Information about Student Performance in Chemistry Examinations? *Research in Science & Technological Education, 31*(1), 49-65.

Johnson, M., & Crisp, V. (2009). An Exploration of the Effect of Pre-Release Examination Materials on Classroom Practice in the UK. *Research in Education, 82*(1), 47-59.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774-1787.

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation, 47*, 93-101.

Kirton, S. B., Al-Ahmad, A., & Fergus, S. (2014). Using Structured Chemistry Examinations (SCHemEs) as an Assessment Method to Improve Undergraduate Students' Generic, Practical, and Laboratory-Based Skills. *Journal of Chemical Education, 91*(5), 648-654.

Lewis, S. E., Shaw, J. L., & Freeman, K. A. (2011). Establishing Open-Ended Assessments: Investigating the Validity of Creative Exercises. *Chemistry Education Research and Practice, 12*(2), 158-166.

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding Their Impact. *Journal of Applied Testing Technology, 13*(3).

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of Automated Scoring of Science Assessments. *Journal of Research in Science Teaching, 53*(2), 215-233.

Moore, R., & Jensen, P. A. (2007). Do Open-Book Exams Impede Long-Term Learning in Introductory Biology Courses? *Journal of College Science Teaching, 36*(7), 46-49.

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300.

Sahin, A. (2013). STEM clubs and science fair competitions: Effects on post-secondary matriculation. *Journal of STEM Education: Innovations and Research, 14*(1), 5.

Seery, M. K., Agustian, H. Y., Doidge, E. D., Kucharski, M. M., O'Connor, H. M., & Price , A. (2017). Developing laboratory skills by incorporating peer-review and digital badges. *Chemistry Education Research and Practice*.

Sikes, S. S., & Schwartz-Bloom, R. D. (2009). Direction discovery. *Biochemistry and Molecular Biology Education, 37*(2), 77-83.

Slepkov, A. D., & Shiell, R. C. (2014). Comparison of Integrated Testlet and Constructed-Response Question Formats. *Physical Review Special Topics - Physics Education Research, 10*(2), 020120-020121-020120-020115.

Stoodley, R., Nunez, J. R. R., & Bartz, T. (2014). Field and In-Lab Determination of Ca2+ in Seawater. *Journal of Chemical Education, 91*(11), 1954-1957.

Tierney, J., Bodek, M., Fredricks, S., Dudkin, E., & Kistler, K. (2014). Using Web-Based Video as an Assessment Tool for Student Performance in Organic Chemistry. *Journal of Chemical Education, 91*(7), 982-986.

Vachliotis, T., Salta, K., Vasiliou, P., & Tzougraki, C. (2011). Exploring Novel Tools for Assessing High School Students' Meaningful Understanding of Organic Reactions. *Journal of Chemical Education, 88*(3), 337-345.

Walpuski, M., Ropohl, M., & Sumfleth, E. (2011). Students' knowledge about chemical reactions - development and analysis of standard-based test items. *Chemistry Education Research and Practice, 12*(2), 174-183.

Appendix 6: Publications identified as potentially relevant to the review but not subsequently included

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: an experimental investigation of focus. Assessment in Education, Vol. 14, No. 2, July 2007, pp. 201-232

Anker-Hansen, J., & Andrée, M. (2015). Affordances and Constraints of Using the Socio-Political Debate for Authentic Summative Assessment. *International Journal of Science Education, 37*(15), 2577-2596.

Barbera, J. (2013). A Psychometric Analysis of the Chemical Concepts Inventory. *Journal of Chemical Education, 90*(5), 546-553.

Bellocchi, A., King, D. T., & Ritchie, S. M. (2016). Context-Based Assessment: Creating Opportunities for Resonance between Classroom Fields and Societal Fields. *International Journal of Science Education, 38*(8), 1304-1342.

Boyce, M. C., & Singh, K. (2008). Student Learning and Evaluation in Analytical Chemistry Using a Problem-Oriented Approach and Portfolio Assessment. *Journal of Chemical Education, 85*(12), 1633-1637.

Broman, K., Bernholt, S., & Parchmann, I. (2015). Analysing Task Design and Students' Responses to Context-Based Problems through Different Analytical Frameworks. *Research in Science & Technological Education, 33*(2), 143-161.

Claesgens, J., Daubenmire, P. L., Scalise, K. M., Balicki, S., Gochyyev, P., & Stacy, A. M. (2014). What Does a Student Know Who Earns a Top Score on the Advanced Placement Chemistry Exam? *Journal of Chemical Education, 91*(4), 472-479.

Cooper, M. M., Sandi-Urena, S., & Stevens, R. (2008). Reliable multi method assessment of metacognition use in chemistry problem solving. *Chemistry Education Research and Practice, 9*(1), 18-24.

Costu, B. (2007). Comparison of Students' Performance on Algorithmic, Conceptual and Graphical Chemistry Gas Problems. *Journal of Science Education and Technology, 16*(5), 379-386.

Cox Jr, C. T., Cooper, M. M., Pease, R., Buchanan, K., Hernandez-Cruz, L., Stevens, R., Holme, T. (2008). Advancements in curriculum and assessment by the use of IMMEX technology in the organic laboratory. *Chemistry Education Research and Practice, 9*(2), 163-168.

Danili, E., & Reid, N. (2005). Assessment formats: do they make a difference? *Chemistry Education Research and Practice, 6*(4), 204-212.

Diegelman-Parente, A. (2011). The Use of Mastery Learning with Competency-Based Grading in an Organic Chemistry Course. *Journal of College Science Teaching, 40*(5), 50-58.

Goubeaud, K. (2010). How Is Science Learning Assessed at the Postsecondary Level? Assessment and Grading Practices in College Biology, Chemistry and Physics. *Journal of Science Education and Technology, 19*(3), 237-245.

Holme, T., Bretz, S. L., Cooper, M., Lewis, J., Paek, P., Pienta, N., . . . Towns, M. (2010). Enhancing the Role of Assessment in Curriculum Reform in Chemistry. *Chemistry Education Research and Practice, 11*(2), 92-97.

Holme, T., & Murphy, K. (2011). Assessing Conceptual and Algorithmic Knowledge in General Chemistry with ACS Exams. *Journal of Chemical Education, 88*(9), 1217-1222.

Hrin, T. N., Milenkovic, D. D., & Segedinac, M. D. (2016). The Effect of Systemic Synthesis Questions [SSynQs] on Students' Performance and Meaningful Learning in Secondary Organic Chemistry Teaching. *International Journal of Science and Mathematics Education, 14*(5), 805-824.

Hudson, R. D. (2012). Is There a Relationship between Chemistry Performance and Question Type, Question Content and Gender? *Science Education International, 23*(1), 56-83.

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics, 89*(3), 337-355.

Jones, I., & Karadeniz, I. (2016, 3-7 August 2016.). *An alternative approach to assessing achievement.* Paper presented at the 40th Conference of the International Group for the Psychology of Mathematics Education, Szeged, Hungary.

Klinger, D. A., & Rogers, W. T. (2006). Differential effects of global modifications to large-scale high stakes examination programmes. *Assessment in Education: Principles, Policy & Practice, 13*(1), 29-43.

Lewis, S. E., Shaw, J. L., & Freeman, K. A. (2010). Creative Exercises in General Chemistry: A Student-Centered Assessment. *Journal of College Science Teaching, 40*(1), 48-53.

Liu, O. L., Lee, H.-S., Hofstetter, C., & Linn, M. C. (2008). Assessing Knowledge Integration in Science: Construct, Measures, and Evidence. *Educational Assessment, 13*(1), 33-55.

Namdar, B., & Shen, J. (2015). Modeling-Oriented Assessment in K-12 Science Education: A synthesis of research from 1980 to 2013 and new directions. *International Journal of Science Education, 37*(7), 993-1023.

Price, P. D., & Kugel, R. W. (2014). The New AP Chemistry Exam: Its Rationale, Content, and Scoring. *Journal of Chemical Education, 91*(9), 1340-1346.

Ruder, S. M., & Straumanis, A. R. (2009). A Method for Writing Open-Ended Curved Arrow Notation Questions for Multiple-Choice Exams and Electronic-Response Systems. *Journal of Chemical Education, 86*(12), 1392-1396.

Schultz, M. (2011). Sustainable Assessment for Large Science Classes: Non-Multiple Choice, Randomised Assignments through a Learning Management System. *Journal of Learning Design, 4*(3), 50-62.

Schwartz, P., & Barbera, J. (2014). Evaluating the Content and Response Process Validity of Data from the Chemical Concepts Inventory. *Journal of Chemical Education, 91*(5), 630-640.

Stanger-Hall, K. F. (2012). Multiple-Choice Exams: An Obstacle for Higher-Level Thinking in Introductory Science Classes. *CBE Life Sciences Education, 11*(3), 294-306.

Vachliotis, T., Salta, K., & Tzougraki, C. (2014). Meaningful Understanding and Systems Thinking in Organic Chemistry: Validating Measurement and Exploring Relationships. *Research in Science Education, 44*(2), 239-266.

Wang, C.-Y. (2015). Exploring General versus Task-Specific Assessments of Metacognition in University Chemistry Students: A Multitrait-Multimethod Analysis. *Research in Science Education, 45*(4), 555-579.

Ye, L., Oueini, R., & Lewis, S. E. (2015). Developing and Implementing an Assessment Technique to Measure Linked Concepts. *Journal of Chemical Education, 92*(11), 1807-1812.

Appendix 7: Glossary: Technical terms used in this review

Dichotomous scoring – occurs where the answer is adjudged to be right or wrong with only two scores possible – normally 1 or 0.

Differentiation – the ability of an assessment to distinguish between students with different abilities

Fixed response questions – questions that require students to select an answer, rather than offer an answer of their own construction

Levels-based mark schemes – describe a number of levels of response, each with an associated mark or band of marks**.**

Closed response questions – answers to these questions are unambiguous, the mark scheme lists acceptable answers.

Construct validity – an assessment has construct validity if the outcomes of the test show a high correlation with other measures of the same construct.

Content validity – an assessment has content validity if the questions and tasks matches the contents and aims of the specification for that assessment, covers the area well, and does not go beyond it.

Inter-rater reliability – an assessment is considered reliable if a script would be awarded the same score if marked by a different examiner using the same mark scheme

Ofqual (Office of Qualifications and Examinations Regulation) – regulates qualifications, examinations, and assessments in England.

Open response questions – (also called constructed or free response) questions that do not constrain the student's response. The score for the questions may be one or two marks or many marks for an essay or other extended piece.

Partial credit – commonly used in scoring open response questions; used to award a proportion of the marks to an answer which is incomplete, or in which an error is made part way through.

Performance assessments – assessment of activities that are models of the real activities a student should be able to carry out, such as a practical investigation or a research report

Points-based mark schemes – provide a list of acceptable points which must be matched by the candidate's answer.

Polytomous scoring – allows scores other than 0 or 1

Tariff – the maximum mark that could be awarded for a particular question.

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

AQA. (2013). AS and A-LEVEL Science in Society 2400. In AQA (Ed.): AQA.

AQA. (2015). *A-level Chemistry (7405/1) Paper 1: Inorganic and Physical Chemistry Specimen 2015 v0.5*: AQA.

AQA. (2016). *GCSE CHEMISTRY Higher Tier Chemistry 1H*: AQA.

AQA. (2017). Projects. Retrieved from http://www.aqa.org.uk/subjects/projects

Baillie, C., & Toohey, S. (1997). The 'Power Test': its impact on student learning in a materials science course for engineering students. *Assessment & Evaluation in Higher Education, 22*(1), 33-48.

Bennett, J. (2003). *Teaching and learning science: A guide to recent research and its applications*. London: Continuum.

Bernholt, S., & Parchmann, I. (2011). Assessing the complexity of students' knowledge in chemistry. *Chemistry Education Research and Practice, 12*(2), 167-173.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning : the SOLO taxonomy (structure of the observed learning outcome) / John B. Biggs, Kevin F. Collis*: Academic Press.

Black, P. (1990). APU science—the past and the future. *School Science Review, 72*(258), 13-28.

Black, P. (1998). *Testing, friend or foe? Theory and practice of assessment and testing*. London: Falmer Press.

Block, R. M. (2012). A Discussion of the Effect of Open-book and Closed-book Exams on Student Achievement in an Introductory Statistics Course. *PRIMUS, 22*(3), 228-238.

Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgement*. Retrieved from Cambridge, UK: http://www.cambridgeassessmentjobs.org/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf

Burgin, S. R., Sadler, T. D., & Koroly, M. J. (2012). High school student participation in scientific research apprenticeships: Variation in and relationships among student experiences and outcomes. *Research in Science Education, 42*(3), 439-467.

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching, 46*(8), 865-883.

Cambridge Assessment. (2017). Thinking Skills Assessment. Retrieved from http://www.admissionstestingservice.org/for-test-takers/thinking-skills-assessment/

Campbell, M. L. (2015). Multiple-Choice Exams and Guessing: Results from a One-Year Study of General Chemistry Tests Designed to Discourage Guessing. *Journal of Chemical Education, 92*(7), 1194-1200.

Charney, J., Hmelo-Silver, C. E., Sofer, W., Neigeborn, L., Coletta, S., & Nemeroff, M. (2007). Cognitive apprenticeship in science through immersion in laboratory practices. *International Journal of Science Education, 29*(2), 195-213.

CIE, (Cambridge International Examinations). (2016). *Cambridge International Level 3 Pre-U Certificate in Global Perspectives and Independent Research 9777*. Cambridge: Cambridge International Examinations.

Cresswell, M. (2000). *Research Studies in Public Examining*. Guildford, Surrey: Associated Examining Board.

Dicks, A. P., Lautens, M., Koroluk, K. J., & Skonieczny, S. (2012). Undergraduate Oral Examinations in a University Organic Chemistry Curriculum. *Journal of Chemical Education, 89*(12), 1506-1510.

Domyancich, J. M. (2014). The Development of Multiple-Choice Items Consistent with the AP Chemistry Curriculum Framework to More Accurately Assess Deeper Understanding. *Journal of Chemical Education, 91*(9), 1347-1351.

Edexcel. (2015a). *Pearson Edexcel Level 3 Advanced GCE in Chemistry (9CH0) Sample Assessment Materials*. London: Pearson Education Limited.

Edexcel. (2015b). Pearson Edexcel Level 3 Extended Project. Retrieved from https://qualifications.pearson.com/en/qualifications/edexcel-project-qualification/level-3.html

Edexcel. (2016). *Pearson Edexcel GCSE (9-1) Chemistry (1CH0) Sample Assessment Materials*. London: Pearson Education Limited.

Eilertsen, T. V., & Valdermo, O. (2000). Open-book assessment: A contribution to improved learning? *Studies in Educational Evaluation, 26*(2), 91-103.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*(6), 915-933.

Fensham, P. J., & Bellocchi, A. (2013). Higher order thinking in chemistry curriculum and its assessment. *Thinking Skills and Creativity, 10*, 250-264.

Francisco, J. S., Nakhleh, M. B., Nurrenbern, S. C., & Miller, M. L. (2002). Assessing Student Understanding of General Chemistry with Concept Mapping. *Journal of Chemical Education, 79*(2), 248.

Grunert, M. L., Raker, J. R., Murphy, K. L., & Holme, T. A. (2013). Polytomous versus Dichotomous Scoring on Multiple-Choice Examinations: Development of a Rubric for Rating Partial Credit. *Journal of Chemical Education, 90*(10), 1310-1315.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying Computerized-Scoring Models of Written Biological Explanations across Courses and Colleges: Prospects and Limitations. *CBE - Life Sciences Education, 10*(4), 379-393.

Hadenfeldt, J. C., Bernholt, S., Liu, X., Neumann, K., & Parchmann, I. (2013). Using Ordered Multiple-Choice Items to Assess Students' Understanding of the Structure and Composition of Matter. *Journal of Chemical Education, 90*(12), 1602-1608.

Hodson, D. (1996). Laboratory work as scientific method: Three decades of confusion and distortion. *Journal of Curriculum Studies, 28*(2), 115-135.

Hudson, R. D., & Treagust, D. F. (2013). Which Form of Assessment Provides the Best Information about Student Performance in Chemistry Examinations? *Research in Science & Technological Education, 31*(1), 49-65.

IB, (International Baccalaureate). (2017). What is the extended essay. Retrieved from http://www.ibo.org/programmes/diploma-programme/curriculum/extended-essay/what-is-the-extended-essay/

Johnson, M., & Crisp, V. (2009). An Exploration of the Effect of Pre-Release Examination Materials on Classroom Practice in the UK. *Research in Education, 82*(1), 47-59.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774-1787.

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation, 47*, 93-101.

Kirton, S. B., Al-Ahmad, A., & Fergus, S. (2014). Using Structured Chemistry Examinations (SCHemEs) as an Assessment Method to Improve Undergraduate Students' Generic, Practical, and Laboratory-Based Skills. *Journal of Chemical Education, 91*(5), 648-654.

Krajcik, J. S., & Blumenfeld, P. C. (2006). Project-based learning. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences.* Cambridge: Cambridge University Press.

Lewis, S. E., Shaw, J. L., & Freeman, K. A. (2011). Establishing Open-Ended Assessments: Investigating the Validity of Creative Exercises. *Chemistry Education Research and Practice, 12*(2), 158-166.

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding Their Impact. *Journal of Applied Testing Technology, 13*(3).

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of Automated Scoring of Science Assessments. *Journal of Research in Science Teaching, 53*(2), 215-233.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and Procedures in TIMSS 2015* Retrieved from http://timss.bc.edu/publications/timss/2015-methods.html

Millar, R. (2004). *The role of practical work in the teaching and learning of science*. Retrieved from Washington, DC:
Moore, R., & Jensen, P. A. (2007). Do Open-Book Exams Impede Long-Term Learning in Introductory Biology Courses? *Journal of College Science Teaching, 36*(7), 46-49.
OCR, (Oxford, Cambridge and RSA Examinations). (2000). *OCR Advanced GCE Chemistry B (Salters) 7887*. Cambridge: OCR.
OCR, (Oxford, Cambridge and RSA Examinations). (2005). *Twenty First Century Science Suite: GCSE Chemistry A J634*. Cambridge: OCR.
OCR, (Oxford, Cambridge and RSA Examinations). (2008). *OCR Advanced GCE Chemistry B (Salters) F336 Chemistry Individual Investigation*. Cambridge: OCR.
OCR, (Oxford, Cambridge and RSA Examinations). (2014a). Extended Project H856. Retrieved from http://www.ocr.org.uk/qualifications/projects-extended-project-h856/
OCR, (Oxford, Cambridge and RSA Examinations). (2014b). *OCR Advanced GCE Chemistry B (Salters) H433*. Cambridge: OCR.
OCR, (Oxford, Cambridge and RSA Examinations). (2016a). *GCSE (9–1) Chemistry A (Gateway Science) J248/03 Paper 3 (Higher Tier) Sample Question Paper*. Cambridge: OCR.
OCR, (Oxford, Cambridge and RSA Examinations). (2016b). *A Level Chemistry B (Salters) H433/01 Fundamentals of chemistry Sample Question Paper*. Cambridge: OCR.
OECD. (2016). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. Paris: OECD Publishing.
Ofqual, (Office of Qualifications and Examinations Regulation). (2014a). *Consultation on the Assessment of Practical Work in GCSE Science* (Ofqual/14/5568 ed.). Coventry.
Ofqual, (Office of Qualifications and Examinations Regulation). (2014b). *Review of Quality of Marking in Exams in A levels, GCSEs and Other Academic Qualifications Interim report* (Crown Ed. Vol. Ofqual/14/5379). Coventry: Ofqual.
Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300.
Sahin, A. (2013). STEM clubs and science fair competitions: Effects on post-secondary matriculation. *Journal of STEM Education: Innovations and Research, 14*(1), 5.
Seery, M. K., Agustian, H. Y., Doidge, E. D., Kucharski, M. M., O'Connor, H. M., & Price , A. (2017). Developing laboratory skills by incorporating peer-review and digital badges. *Chemistry Education Research and Practice*.
Sikes, S. S., & Schwartz-Bloom, R. D. (2009). Direction discovery. *Biochemistry and Molecular Biology Education, 37*(2), 77-83.
Slepkov, A. D., & Shiell, R. C. (2014). Comparison of Integrated Testlet and Constructed-Response Question Formats. *Physical Review Special Topics - Physics Education Research, 10*(2).
SQA, (Scottish Qualifications Authority). (2010). Chemistry Open-ended Questions Support Materials. Retrieved from http://new.chemistry-teaching-resources.com/Resources/Higher/LTS/OpenEndedQuestions_tcm4-628995.pdf
SQA, (Scottish Qualifications Authority). (2016). Chemistry Assignment General assessment information. Retrieved from http://www.sqa.org.uk/files_ccc/GAInfoHigherChemistry.pdf
Stoodley, R., Nunez, J. R. R., & Bartz, T. (2014). Field and In-Lab Determination of Ca2+ in Seawater. *Journal of Chemical Education, 91*(11), 1954-1957.
Thurstone, L. L. (1959). The measurement of paired comparisons for social values *The measurement of values*. Chicago, Illinois: University of Chicago Press.
Tierney, J., Bodek, M., Fredricks, S., Dudkin, E., & Kistler, K. (2014). Using Web-Based Video as an Assessment Tool for Student Performance in Organic Chemistry. *Journal of Chemical Education, 91*(7), 982-986.
University of Cambridge. (2017). Natural Sciences Admissions Assessment Retrieved from http://www.undergraduate.study.cam.ac.uk/courses/natural-sciences#entry-requirements
Vachliotis, T., Salta, K., Vasiliou, P., & Tzougraki, C. (2011). Exploring Novel Tools for Assessing High School Students' Meaningful Understanding of Organic Reactions. *Journal of Chemical Education, 88*(3), 337-345.

Walpuski, M., Ropohl, M., & Sumfleth, E. (2011). Students' knowledge about chemical reactions - development and analysis of standard-based test items. *Chemistry Education Research and Practice, 12*(2), 174-183.

Wertheim, J., Osborne, J., Quinn, H., Pecheone, R., Schultz, S., Holthuis, N., & Martin, P. (2016). *An analysis of existing science assessments and the implications for developing assessment tasks for the NGSS*. Retrieved from https://web.stanford.edu/group/ngss_assessment/cgi-bin/snapgse/?page_id=193

Whitehouse, M. (2014). *Developing and testing a theoretical framework for assessing extended response questions in GCSE Science.* (MA by research), University of York, York. Retrieved from http://etheses.whiterose.ac.uk/id/eprint/6197