Cambridge Assessment

Research *Matters*

Issue 27 / *Spring 2019*

Cambridge Assessment

Proud to be part of the University of Cambridge

Established over 150 years ago, Cambridge Assessment operates and manages the University's three exam boards and carries out leading-edge academic and operational research on assessment in education. We are a not-for-profit organisation.

Citation

Articles in this publication should be cited using the following example for article 1: Bramley, T., Crisp, V., & Shaw, S. (2019). The art of test construction: Can you make a good Physics exam by selecting questions from a bank? *Research Matters: A Cambridge Assessment publication*, *27*, 2–9.

Credits

Reviewers: Tom Benton, Ellie Darlington, Sarah Mattey, Sylvia Vitello and Joanna Williamson, Research Division, Cambridge Assessment
Editorial and production management: Karen Barden, Research Division, Cambridge Assessment
Additional proofreading: David Beauchamp, Research Division, Cambridge Assessment
Cover image: John Foxx Images
Design: George Hammond
Print management: Canon Business Services

All details are correct at the time of going to print in May 2019

If you would like to comment on any of the articles in this issue, please contact Tom Bramley, Director, Research Division. Email: researchprogrammes@cambridgeassessment.org.uk

This and all previous issues of *Research Matters* are available to download, in full and by individual article, from our website: www.cambridgeassessment.org.uk/research-matters

Research *Matters* / 27

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

There currently is a great deal of interest in the impact of the new model of assessment in GCSE and GCE Science – where performance in practical work does not contribute directly to the grade – see the related article in this issue by Frances Wilson and colleagues. The new arrangements initially were monitored in pilot schools, and are now, rightly, the focus of considerable evaluation and review. The model was designed carefully – to assess the knowledge and understanding which is essential to practical work in science. And the model tries to break "assessment dominated learning", by promoting and emphasising the requirement to undertake a rich diet of practical as part of learning programmes. Some critics stated it would "kill" practical work; current evidence suggests the opposite. But where did the model come from? Before the mists of time descend on the history, it might be useful to trace things.

The bidding process for delivery of *English Baccalaureate Certificates* may be forgotten, but in 2013 exam boards hunkered down to prepare competitive bids for GCSE replacements. In the wake of work by OCR colleagues in 2005 on an alternative to practicals, I looked at what practicals were actually for – the work of Robin Millar, then Board Director, National Skills Academy UK, was particularly valuable – the problematic claims that practicals give access to "real science", the enduring problems of controlled assessment, and the professional contradictions being placed on teachers. The latter were particularly important. On the one hand we were asking teachers to be independent agents of exam boards (in making consistent assessment contributing to the grade) and on the other, there was huge pressure on teachers to gain maximum marks in the practical assessment, and improve grades year on year, on behalf of the school. Tightening the conditions of the assessment had proved dysfunctional. So, on 12 December 2012, I recommended to Simon Lebus (CEO, Cambridge Assessment, 2012–2018) a controversial approach which I believed would be an essential and unique part of OCR's bids; taking practicals out of formal assessment. This proposal was refined and developed into a working model by colleagues in OCR. The rest is public record, and the model moved from being part of the abandoned bidding process to being national policy, where it seems to have achieved its objectives – focused and dependable examinations, and rich practical work in schools.

Tim Oates, CBE *Group Director, Assessment Research and Development*

Editorial

Two of the articles in this issue of *Research Matters* are related to the impact of technology on assessment. One area where there is potential for technology to change the way we do things is in how exams are constructed. In GCSEs and A levels, the traditional way is for one person to write all the questions. But in many other assessments these days, the test is constructed by selecting questions from a bank of questions. In the first article, Vicki Crisp, Stuart Shaw and I report on how the method of construction affects expert perceptions of the quality of the resulting exam paper.

A second area that all readers will be aware of is how increases in computing speed and power, and data storage and transmission capability, can transform the kinds of knowledge we can gather and inferences we can make about what students know and can do. In the third article, we publish an edited transcript of a conference presentation by Nicholas Raikes that guides us through the hype and identifies some of the opportunities (and dangers) that are currently being explored in the assessment world.

The fourth article by Lucy Chambers, Joanna Williamson and Simon Child describes what we believe is the first detailed qualitative study of the cognitive processes involved in moderating Art and Design – would the processes found for more traditional assessments also apply to the very different type of work produced here?

Our final article is something of a departure for *Research Matters*. While we have often reported our attempts to use insights from research in psychology to improve how we develop, mark and grade assessments, here Gill Elliott, Irenka Suto and Emma Walland turn their attention to what we can learn from psychology to make us more confident and effective at sharing that research when we have to stand up and talk about it – public speaking being no less stressful for researchers than for anyone else!

Tom Bramley *Director, Research Division*

The art of test construction: Can you make a good Physics exam by selecting questions from a bank?

Tom Bramley, **Victoria Crisp** Research Division, and **Stuart Shaw** Cambridge Assessment International Education

Introduction

The traditional approach to producing an examination paper of the type found in General Certificate of Secondary Education (GCSE) and General Certificate of Education Advanced Level (GCE A Level) assessments has been for a single person – a subject matter expert and usually a former or practising teacher – to write the whole paper. They write each question so as to ensure that the topics and assessment objectives set out in the syllabus are suitably well covered, and that the questions are appropriately targeted at the examinees in the range of ability for which the exam is intended. A variety of individuals and committees are involved in the many activities and checks that make up the question paper production process as a whole, but it is nearly always still the case that a single mind is behind the set of questions that eventually appears in the paper on the desk in the examination hall. This traditional approach to exam paper construction could be given the label "creating".

The technological advances of recent decades have led to innovations and developments in assessment, most obviously the arrival of computer-based testing. For many types of assessment (though not GCSEs and A Levels) it is now routine for examinees to take the test on a computer. Often these tests are available on demand, and some are adaptive (in the sense that the next question presented to an examinee depends on their success on previous questions). In most of these instances, the tests are constructed by selecting the questions from a bank of suitable questions. This selection can be done either by humans or by computer (in the case of adaptive testing it is by computer). The bank of questions will usually be large and will contain questions created by numerous authors. The particular combination of questions presented to an examinee has a "mind behind it" when the questions have been selected from the bank by an individual or team, and no mind behind it at all if selected by a computer (unless in the sense that the algorithm for selecting the questions will have been created by humans). This approach to exam paper construction could be given the label "compiling". Note that whilst this compiling approach is often used for computer-based tests, it can also be used where the test will be paper based.

There are many good reasons why the compiling approach is not yet commonplace for GCSEs and A Levels, including the large number of questions that are needed in the bank to allow the test constructor to meet all the constraints imposed by the specification (i.e., balance of topics, skills and difficulty). A significant further obstacle is that in most GCSE and A Level examinations, the questions are permitted to vary (sometimes substantially) in the number of marks they are worth. Thus, a test constructor of a Biology exam might find themselves needing to locate a 7-mark question testing knowledge of respiration with the further restriction that it should not contain a graph if graph-interpretation skills have already been assessed in other questions selected thus far. Clearly the bank of questions needs to be very large to give them a reasonable chance of finding a suitable question. In the discussion, we consider some ways in which the test construction process could change to facilitate a compiling approach.

Whether a single creative mind needs to be behind the full set of questions, to ensure that they cohere and achieve an appropriate balance of content and skills, is currently unclear. From various informal conversations with professionals involved in the question paper production process, we gained the impression that they felt a compiling process would be detrimental to quality for typical GCSE and A Level papers. We carried out a two-stage study to investigate issues relating to compiling an examination paper from an item bank. The first stage, reported in Crisp, Shaw and Bramley (2018), was a detailed investigation of the issues faced by test constructors when compiling a paper. The second stage, reported here, was an evaluation of the perceived quality of exam papers constructed by different methods. We wanted to test whether in fact assessment experts could distinguish between tests that had been created and compiled, when they were unaware of the method of construction.

Method

Exam papers

Seven Physics General Certificate of Education Advanced Subsidiary Level (GCE AS Level) exam papers were used in order to investigate experts' views on papers constructed in different ways. Two of the papers were actual past exam papers, created in the usual way. Three papers had been constructed (compiled) by subject experts in the first stage of the study from a bank of 175 questions that had been used on past exam papers (see Crisp et al., 2018, for details of the bank and the construction process), and two were constructed semi-automatically using an algorithm – more details follow in the article. We thought it would be interesting to include papers that had been constructed automatically because, whilst experts might believe that it is necessary to have a fine balance of various quality-related features (not all of which can be quantified and coded) in order to make a 'good' paper, if they were not able to distinguish between the computer-compiled ones and the expert-compiled ones in terms of quality, this would weaken the idea that test construction is an "art" that can only be carried out by an expert.

The exam used was an international AS Level Physics paper, out of 60 marks in total, generally comprising around 6 structured questions (made up of part questions) worth around 6 to 11 marks. In the normal test creation process, the question paper setter completes a specification grid or "setting grid" recording which syllabus topics and subtopics are tested in each part-question, how many marks are

assigned to the two Assessment Objectives (AOs) – which also have numbered subdivisions, and how many marks are assigned to different ability levels or "target grades" (A/B, C/D, and E/U). There are some constraints that must be met in terms of how the marks are allocated: for this particular paper the weightings of the AOs are mandated by a statement in the syllabus that the balance in Paper 2 will be approximately 48% from AO A (Knowledge with understanding) to 52% from AO B (Handling, applying and evaluating information), which gives an ideal target of 29 marks on AO A and 31 marks on AO B. However, this is stated as the approximate weighting, and since the setting grids from past papers revealed a range of 25 to 29 marks for AO A, we used this range for our study. There are no officially mandated targets for the number of marks targeted at each grade band. However, discussion with question writers and staff involved in the normal production of this paper suggested that there were approximate targets based on discussions between them which had become established practice. We therefore used both the official and unofficial established constraints when creating our algorithm for the automatic compilation.

Writing an algorithm to construct papers that would meet all the relevant criteria would have been difficult and time-consuming (if it were possible at all), but it was relatively easy to write an algorithm to construct papers worth 60 marks by selecting whole questions from the bank. The two semi-automatically generated papers used in the study were created as follows:

- 500 60-mark tests were created by sampling whole questions from the bank.
- From these, the tests where every question tested a different main topic[1] were retained.
- From these, the tests that met the following four targets were retained:
    1. Number of marks for AO A between 25 and 29 (and hence the number of marks for AO B between 31 and 35).
    2. Number of marks targeting grades A and B between 17 and 20.
    3. Number of marks targeting grades C and D between 22 and 25.
    4. Number of marks targeting grades E and U between 17 and 20.

A total of 9 tests from the original 500 met all 5 targets and were retained. At this point, there was human intervention to get to the final two tests. We checked to see whether the secondary topic overlapped with the main topic on different questions (which would have created less wide-ranging, and possibly repetitive, papers) and selected the best two papers in terms of breadth of main and secondary topics. Finally, we read through the papers to check that there was nothing that would make it glaringly obvious that the test had been constructed by computer. We found one instance of the same subtopic (the Young modulus) appearing as part of two different whole questions on the same paper. Whilst the questions did test different skills, it seemed unlikely that both would in practice appear on one paper. We therefore replaced one of the whole questions with a different question testing the same main topic and worth the same number of marks. The resulting two computer-generated tests were therefore not wholly automatically generated, but neither were they generated by Physics experts. It was easy to decide on the order of questions for the computer-generated papers because the practice for this particular paper is to put the questions in syllabus order by topic. Therefore the ordering could be done automatically.

Using Portable Document Formats (PDFs) of the individual questions from past papers which comprised the bank, a new PDF for each of the seven papers was created. The questions were numbered into order, and a cover page and page numbering were added so that the real papers looked no different from the expert-compiled and computer-compiled papers. Mark schemes were created for the papers in the same way, and setting grids were compiled in a consistent format. The seven papers were randomly assigned letter codes (H to N) to identify them.

Procedure

Three experts were involved, all with experience of reviewing and/or setting Physics exam papers at AS and A Level. Two of them had been involved in the test construction stage of the research study. They conducted the evaluation task at home. We asked them each to evaluate six of the question papers, as follows:

- Two of the three papers compiled by participants in the test construction stage of this research study (not papers that they themselves had compiled if they were involved).
- The two actual past papers.
- The two papers compiled semi-automatically by computer.

We did not give participants who had also taken part in the test construction stage their own papers to evaluate because, if they recognised their paper, this could have influenced their reactions. But one paper from each participant in the construction stage was evaluated by two participants in the evaluation stage. Thus, seven papers in total were involved. We did not tell the participants that the papers had been constructed in different ways.

We decided to collect the participants' evaluations of the papers in two parts. This was because we did not want to ask leading questions that might draw their attention to features of the papers that they would not otherwise have paid attention to, and we did not want to assume that they all defined question paper quality in the same way. The first part of the evaluation was therefore more open-ended. They were initially asked to define "quality" as it applies to a question paper. They were then asked the same set of questions about each of the six papers they were asked to consider. These questions were aimed at finding out how far short from the ideal the paper fell: first, in terms of the number of whole questions that would need to be replaced for it to be useable (which is what would need to happen if the only way papers could be constructed was by assembling whole questions from the bank); and second, in terms of whether an acceptable paper could be created by editing subparts of the existing questions (which is what could happen if the role of the item bank were more that of a "set of resources" in the test construction process). Participants were asked to provide reasons for these evaluations, including strengths and weaknesses of the papers. Once they had completed the first part of the evaluation and sent their responses back to us, we sent the participants the evaluation questionnaire for the second part. This was more closed – they were asked a set of specific questions about each of the papers. These specific questions reflected our concerns, and those of experts we had spoken to, about the potential pitfalls of creating tests by selecting questions from

1. For each whole question, we defined the main topic to be the one with most marks coded against it on the setting grid, and the secondary topic to be the one with the second most marks coded against it across all the subparts of the question.

a bank. The concerns covered: balance of AOs, topics and target grades; incline of difficulty; repetition of topics or skills; and instances of parts of one question giving away the answer to parts of different questions. The final question asked participants if they noticed anything odd, unusual, out of character, or inappropriate about the paper. This question was asked as a way of discovering whether the computer-generated tests stood out to the participants as being different.

Results

Questionnaire: Part 1

The first question asked participants to define "quality" as it relates to an exam paper. The participants' responses are summarised below.

Features of quality relating to the paper as a whole:

- *Range of question types avoiding repetition of same skill/process.*
- *Good coverage of syllabus (in conjunction with the other components of the examination[2]).*
- *Correct balance of the two AOs, with most questions having elements of both.*
- *Can be completed in the time available and can't be completed too quickly by the best candidates.*
- *Should differentiate well (produce a good spread of marks in the target cohort).*
- *Should challenge candidates of all abilities.*
- *Should meet criteria of the vetter's checklist (e.g., sufficient space to write answers, not radically different from previous papers, does not disadvantage particular groups, etc.).*
- *Should flow well with a logical order of topics.*
- *Should be reliable.*

Features of quality relating to the individual questions:

- *Questions should be clearly written and unambiguous.*
- *All parts of all questions should be accessible to the candidates.*
- *The context of questions should be realistic and, ideally, original and interesting.*

The second question aimed to establish whether the participants felt that the papers were good enough to be used and, if not, how much change was needed. Table 1 shows the participants' responses by paper.

Table 1: Summary of evaluation of the seven papers

Note: The three participants' responses are recorded (in the same order) for each paper.

Source of paper	*Paper ID*	*Good enough to be used?*	*Needs one whole question replaced?*	*Needs two or more whole questions replaced?*	*Would be OK if I could edit subparts?*
Actual (created)	I L	NNN NNN	NNY YNN	YNN PNY	#Y# YY#
Expert compiled	H K M	YN- N-N -NN	PY- Y-N -NN	NN- N-Y -YY	##- Y-# -##
Computer compiled	J N	NNN NNN	YYY NNN	NNN Y#Y	Y## #YY

Key: Y = Yes; N = No; P = Possibly; # = No response; - = Not asked.

It seems from Table 1 that the participants in this study were quite harsh critics of exam papers! Only one of the seven papers was deemed to be good enough to be used, and that was by just one of the three participants. Computer-compiled *Paper N* and expert-compiled *Paper M* were clearly considered to be the worst, with unanimous agreement that they would need either two or more whole questions to be replaced, or editing of the subparts. The actual past papers fared little better, with two out of the three participants thinking they would need either two or more whole questions to be replaced, or editing of the subparts. The other participant in each case felt that one whole question needed to be replaced. Computer-generated *Paper J* and the expert-constructed *Papers H* and *K* seemed to be the best, in general being deemed to require only one whole question to be replaced, or to need editing of subparts.

However, examining the open-ended responses about the reasons for these evaluations, the picture is not quite so clear cut. Tables 2 to 4 summarise the participants' descriptions of the strengths and

2. A multiple-choice paper and an assessment of practical skills.

Table 2: Actual past papers – evaluation of strengths and weaknesses, by participant

Paper I

Strengths	*Weaknesses*
Adequate differentiation AOs well balanced, but many calculations and few explanations required Most question parts accessible to average candidate	Paper is not well balanced – similar areas of syllabus tested and two important areas (4 & 5) not covered in depth Overlap in testing resolution of vectors and energy in Q2 and Q3
Good range of key topics Appropriate level of difficulty Some tricky calculations which will differentiate Good balance of recall versus application	Some formatting issues – but could just be errors when compiling these sample papers
Starts with a good, accessible question to settle nerves Some more challenging descriptive parts Diagrams and graphs to interpret and draw information from Overall, this is a good paper	Overemphasis on mechanics (Topics 1–6, 9) Nothing on Topics 17 or 26 Overemphasis on Skill A1, with little on other AO A skills (though these can be hard to test) There could be a question to test AO B4 (Trends and patterns)

Paper L

Strengths	*Weaknesses*
Good balance of learning outcomes Good variety of type of question Some good questions	Overemphasis on AO B Underemphasis on E/U marks Paper may be slightly on difficult side (complex topics and few "easy" marks) Some challenging parts
Good mix for AO A and AO B Appropriate level of difficulty Covers most topics Some challenging elements Some good contexts	No obvious weaknesses
Q1 is easy access for all candidates Candidates draw a vector triangle (as well as a graph)	Too few AO A marks Too many C/D marks Nothing on Topics 2 or 14 No graphs or diagrams to read or interpret Q1 and Q2 set in a similar context Limited range of skills within AO A and AO B

Table 3: Expert-compiled papers – evaluation of strengths and weaknesses, by participant

Paper H

Strengths	*Weaknesses*
Balance of AOs Good differentiation Reasonable syllabus coverage Variety of question types	Too many difficult parts
Covers many of major topics Good balance between recall and application Easier and more difficult elements to most questions No obvious duplication of material	Q1 could be extended Q4 disjointed

Paper K

Strengths	*Weaknesses*
Good balance for AOs Good variety of questions Good questions	Too many harder topics (e.g., momentum) Too many difficult parts (but grid doesn't reflect this) Q1 and Q2 set in similar contexts (ball falling) Questions not in syllabus order
Candidates need to gather information from a graph and interpret a diagram/graph	Key topics missing (1, 2 & 4) Questions not in logical syllabus order First question too difficult Q1 and Q2 set in similar contexts Overemphasis on descriptive work compared to calculation

Paper M

Strengths	*Weaknesses*
Most questions test AO A and AO B Most questions have simpler and harder parts Should result in a good range of marks	Predictable "textbook" contexts (e.g., car travelling on road, waves in a ripple tank) – not very interesting
Candidates describe trend in a graph and give reasons Easy first question	Some topics omitted (2, 5, 6 & 20) Overemphasis on two topics (7 marks on base units, 19 marks on waves) Underemphasis on mechanics (3, 4, 5, 6 & 9) Not enough on graph/diagram skills where candidates interpret or draw their own

Table 4: Computer-compiled papers – evaluation of strengths and weaknesses, by participant

Paper J	
Strengths	*Weaknesses*
Good balance of AOs Good differentiation Good coverage of topics	Overlaps in topics (potential energy, power) Nothing on Topic 4 Some parts too difficult
Good balance between explanations and application Most questions well structured	Mostly predictable contexts (e.g., output power of an electrical heater)
Graph that needs to be read/interpreted and a table to complete Overall, a good paper	Nothing on Topics 14 and 20 Overemphasis on Skill A1 with little on other AO A skills (though these can be hard to test)
Paper N	
Strengths	*Weaknesses*
Good starter question Good balance of AOs Reasonable differentiation Good accessibility in majority of questions	Overlaps in concepts (e.g., Q3, Q4 and Q6 relate to equilibrium of forces and Newton's second law, – mass x gravitational field strength calculated in each of these questions)
Good range of topics Q4 particularly good – good context, both AO A and AO B marks, and combines two topic areas Some difficult questions to test more able candidates	No obvious weaknesses
Good coverage of most of syllabus Graph drawing accuracy is tested Balance of setting grid looks OK	Nothing on one key topic (3) Overlaps in topics (Topic 9, Subtopic 4.2a) No graphs or diagrams to interpret or gather information from Too many easy marks Too few A/B marks on Q7

weaknesses of each paper. In each table, the first row summarises the responses of *Participant 1*, the second those of *Participant 2*, and so on.

It seems from these comments that all of the papers were in fact evaluated less harshly than the overall judgements in Table 1 might have suggested. Some of the reasons given for why the paper had not been deemed usable related to concerns about specific questions, rather than features of the paper as a whole. The particular concerns of the different participants were also apparent – one made far more comments about the details of individual questions than the other two; another referred several times (in 'Other comments') to not being able to assess how long it would take examinees to complete the papers without attempting the questions themselves.

Overall, the range of comments does not suggest that papers compiled by selecting whole questions from a bank are necessarily worse (or better) than those created in the usual way. However, they do highlight how difficult it is to create papers that satisfy all the constraints, and meet all the criteria for quality that experts in assessing Physics aim to achieve.

Questionnaire: Part 2

As described earlier, in the second part of the evaluation work, the participants were asked a number of more specific questions about each of the papers they evaluated. Their responses are shown in Table 5.

Table 5 shows, again, that the different participants had consistently different views about some papers. For example, *Participant 3* was more likely to agree there was a general increase in difficulty, but less likely to agree that there was an appropriate balance of AOs or target grades. *Participant 2* tended to note repetition of skills (such as substituting numbers into formulas), whereas the other two did not. *Participant 2* was also more likely to pick up on odd, unusual, out of character, or inappropriate features of papers. These often related to features of individual questions, rather than anything about how the questions combined together. It is interesting to note that the computer-compiled papers (especially *Paper N*) were more likely to be judged to have repetition of learning outcomes and skills than the other papers. This is likely to be because there were many aspects that the automatic construction process ignored, such as the number of marks allocated to secondary topics, and the finer-grained categories of the AOs. One of the questions where one part was deemed to give away the answer to another part was within the same question, so it was not an issue of compiling questions. The other (on one of the expert-compiled papers) arose because there were two graph questions where the shape of one graph would have hinted at the correct shape for the other.

Discussion

When asked to evaluate a number of papers (some actual past papers and some created from the bank by the participants or by computer), experts identified ways in which all papers fell short of the ideal, some more than others. There were no consistent patterns relating to how each paper had been constructed, which indicates that the papers constructed from the bank by a compiling process were not inherently worse than papers created by the usual method. The participants defined

Table 5: Summary of closed-question evaluation of papers

	Actual (created)		*Expert compiled*			*Computer compiled*	
	I	*L*	*H*	*K*	*M*	*J*	*N*
Is there a general increase in difficulty through the paper?	NNY	NNY	NN-	N-N	-NY	NYY	NYY
According to the setting grid, this paper meets the targets for the balance of Assessment Objectives. Looking at the paper, do you feel that the balance is appropriate?	NYN	NYN	YY-	Y-N	-YN	YYN	YYN
According to the setting grid, this paper meets the targets for the balance of target grades. Looking at the paper, do you feel that the balance is appropriate?	YYY	NYN	YY-	N-N	-YN	NYY	YYN
Is there a suitable balance of learning outcomes?	NYN	YYY	YY-	Y-Y	-YN	YYN	NYN
Is there any repetition of learning outcomes in the paper?	YNN	NNN	NN-	N-N	-NN	YNN	YNY
Is there any repetition of skills in the paper (e.g., graph work, a particular type of calculation)?	NYN	NYN	NY-	N-N	-YN	NYN	NYY
Do any questions give away parts of an answer to another question?	NYN	NNN	NN-	N-Y	-NN	NNN	NNN
Have all key topics that should be included in all papers been included?	NNY	YYN	N?-	Y-N	-YY	NNN	NYN
Is there anything odd, unusual, out of character, or inappropriate about this paper? If so, please specify.	NYN	NNY	NY-	Y-N	-YY	NYN	NNN

Key: Y = Yes; N = No; ? = Not sure; - = Not asked.

quality in exam papers as might have been expected, (i.e., relating to themes such as coverage of the syllabus and AOs, differentiation, being achievable in the time available, and including a range of question types requiring different skills/processes). However, when they were evaluating the papers for quality, they often focused on characteristics of individual questions rather than characteristics of the test as a whole. Compiling tests semi-automatically by computer algorithm followed by non-expert review and tweaking produced one test that was rated relatively well, and one that was rated relatively badly, so we have not learned enough from this experiment to be able to recommend using or avoiding semi-automatic compilation of this kind of question paper from a bank.

In the remainder of this discussion we attempt to relate the findings of this study to the wider context of item banking of structured questions. Test construction from an item bank could be characterised as a constraint satisfaction problem[3] where a solution needs to be found within certain imposed constraints or conditions. Such problems arise in a very wide variety of areas. In the particular case studied here, the target was to compile a paper worth a total of 60 marks, subject to the following constraints:

- Questions must only cover topics that are on the (AS) syllabus.
- Topic coverage must fit with (i.e., complement rather than repeat) topic coverage on other components of the examination.
- Questions must not be reused.
- *The paper total must equal 60 marks.
- *Each whole question should test a different main topic.
- *25–29 marks should test AO A and 31–35 marks should test AO B.
- *17–20 marks should target grades A/B.
- *22–25 marks should target grades C/D.
- *17–20 marks should target grades E/U.
- All the topics on the syllabus should be covered over a period of *x* years.
- Every paper should test at least *n* of the following *m* 'key topics'.
- Within the marks allocated to each AO, there should be a good balance of the AO subcategories.
- There should be a variety of contexts across the questions in the paper.
- One question or question part should not give away the answer to another question or question part.

* These constraints were the ones we applied in our computer generation method.

Most of the constraints we have listed clearly relate to the definitions or characteristics of quality provided by the experts. However, their judgements were expressed in qualitative terms and it may be that the attempt to quantify them by assigning specific mark allocations on the setting grid is too constraining. In the question paper used in our study, the constraints for the number of marks testing each AO and target grade had ranges rather than specific values, recognising first that it might be difficult to meet exact targets (even if constructing a paper the traditional way), and second that there may be subjectivity (room for expert disagreement) on how to allocate marks to AOs and target grades (see Crisp et al., 2018). But is there evidence showing that these constraints, and the particular values they take, contribute to assessment quality? Further research could perhaps ask experts to judge the qualities of constructed papers that did not meet these constraints. It is certainly worth questioning whether the constraint on marks at target grades is worthwhile, given that it is difficult to define coherently what is meant by a "mark targeting a grade", and that expert judgement of item difficulty often does not correlate particularly well with actual difficulty in terms of the marks gained by examinees (e.g., Bejar, 1983; Brandon, 2004). Further research could explore whether assigning target grades does actually help with standards maintenance. The setting grid and allocation of grade targets within it also potentially serve as an accountability function of recording that thought has been put into checking that a paper includes questions ranging in difficulty. However, it is possible that the accountability function could be maintained, and that the standard maintaining function could be improved, if a different kind of judgement about question difficulty was collected – namely the expected mean score that candidates on a key boundary would obtain. See Bramley and Wilson (2016) for full details.

3. See, for example, https://en.wikipedia.org/wiki/Constraint_satisfaction

Clearly the more constraints there are, the more difficult it is to satisfy them all. In this particular context, the ease of meeting the constraints clearly depends on the size and variety of the item bank (including the nature and range of questions in it). By analogy, if the task were to spend exactly £60 on food with constraints on the proportion spent on mutually exclusive categories such as meat, fruit, vegetables, dairy products, and so on, and with other constraints on categories cutting across these categories such as frozen or non-frozen, and so on, it would probably be easier to achieve the task in a supermarket than a corner shop due to a greater variety of products being available. The first stage of this research (Crisp et al., 2018) had shown that, even with around 20 times as many questions in the bank as needed for a single paper, experts still found it difficult to compile a 60-mark paper meeting the constraints. The main contributory factor to this difficulty is that, traditionally, most questions in GCSE and A Levels are allowed to vary in how many marks they are worth. It would therefore be sensible, if the test construction process were to change from being one of creating to one of compiling, to stipulate a standard set of mark tariffs for questions. For example, if Physics exam questions were limited to tariffs of 1, 2, 5 and 10 marks, and the test compilation process specified the combination needed for the overall paper (e.g., 3 x 10-mark questions, 4 x 5-mark questions, 3 x 2-mark questions, and 4 x 1-mark questions) then the bank would not need to be so large as it would if questions could be worth any tariff. Furthermore, the bank could be built up intelligently by commissioning questions at the different tariffs in the proportions needed to allow construction of high-quality papers by a compiling process. An initial reaction from question-writers to such a suggestion might be that constraining mark tariffs would reduce flexibility and, therefore, reduce question quality. However, there is no evidence available to inform us on whether this would actually be the case. It may be that there is a kind of circularity in effect, whereby writers need flexibility to vary the numbers of marks they can assign to individual questions in order to meet constraints on the setting grid for mark allocations at whole paper level (Bramley, 2001). However, it is worth noting that question writers (at least in some subjects and with some types of questions) are quite capable of writing questions worth the same mark total because this is necessary whenever exam papers contain sections where questions are optional, as used to be the case in General Certificate of Education Ordinary Level (GCE O Level) Physics (Bramley & Crisp, 2018). Further research is needed to see whether imposing more rigid constraints on question tariffs would have a negative effect on question quality. One factor that might need to be taken explicitly into consideration is linking the mark tariff to the time it would take to answer the question, in order to ensure that papers with the same total mark could be completed in the same amount of time.

In conclusion, we have not found strong evidence that question papers that are compiled are of different quality (as perceived by experts) to those that are created. While we might be reasonably confident that the findings from this study would generalise to subjects with similar types of questions and constraints in the test construction process, future research could consider subjects with different types of questions and constraints. If compilation were to become the normal process for constructing papers of this type, however, it may be necessary to rethink some of the flexibilities and constraints found in the traditional creating process.

References

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*(3), 303–310.

Bramley, T. (2001). The question tariff problem in GCSE Mathematics. *Evaluation and Research in Education, 15*(2), 95–107.

Bramley, T., & Wilson, F. (2016). Maintaining test standards by expert judgement of item difficulty. *Research Matters: A Cambridge Assessment publication, 21*, 48–54.

Bramley, T., & Crisp, V. (2018, November, 29). *Spoilt for choice? Is it a good idea to let students choose which questions they answer in an exam?* [Blog]. Retrieved from http://www.cambridgeassessment.org.uk/insights/spoilt-for-choice-is-it-a-good-idea-to-let-students-choose-which-questions-they-answer-in-an-exam/

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*(1), 59–88.

Crisp, V., Bramley, T., & Shaw, S. (2018, November 7). *Should we be banking on it? Exploring potential issues in the use of 'item' banking with structured examination questions.* Paper presented at the 19th Annual Conference of the Association for Educational Assessment in Europe, Arnhem-Nijmegen, The Netherlands.

Indirect assessment of practical science skills: Development and application of a taxonomy of written questions about practical work

Frances Wilson OCR, **Stuart Shaw** Cambridge Assessment International Education,
Neil Wade OCR, **Sarah Hughes** OCR, and **Sarah Mattey** Research Division
(The study was completed when the fourth and fifth authors were based in Cambridge Assessment International Education)

Background

Practical work is central to science education and is used not only to support the development of conceptual knowledge, but to enable students to develop a wide range of skills, including data handling, experimental design and equipment manipulation (Wilson, Wade, & Evans, 2016). Practical work may be assessed using many different forms of assessment, including coursework projects, practical exams and written questions in exams. Different forms of assessment may assess different aspects of this complex domain. As such, it is important to establish a clear understanding of the skills and knowledge which are assessed by each form of assessment. In this article, we focus on the evaluation of one method for the assessment of practical science: written exam questions.

Abrahams and Reiss (2012) define practical work as "an overarching term that refers to *any* type of science teaching and learning activity in which students, either working individually or in small groups, are involved in manipulating and/or observing real objects and materials" (p.1036). This is as opposed to virtual materials, such as those obtained from a video or digital simulation, or even from a text-based account. Furthermore, Lunetta, Hofstein, and Clough (2007, p.394), suggest that practical activities may:

1. be experienced in school settings where students interact with materials to observe and understand the natural world;
2. be individual or group, or large group demonstrations;
3. vary on a continuum between highly structured and teacher-centred through to open inquiry;
4. last several weeks, including outside school activities, or less than 20 minutes;
5. use a high level of instrumentation or none at all; and
6. include activities where students gather data to illustrate a principle/relationship (deductive), and those where students gather data and try to work out patterns or relationships from the data (inductive).

Practical work may therefore span a wide range of different activities, which may be used by teachers to serve many different learning aims. Given the impact of assessment on what is taught and learned, it is important to understand exactly what skills are assessed, and how they relate to the specified aims of practical work within the curriculum. Previous work in this area has articulated multiple purposes for practical science, in varying levels of granularity, and based on different interpretations of what is considered to be practical work. For example, focusing primarily on higher education (HE), Kirschner and Meester (1988) catalogued 120 different aims for practical work, which they synthesised into 8 overarching aims. However, these aims are not given equal importance by teachers. Additionally, the relative importance of the aims of practical work at different stages of education and across the different science subjects may vary (Kerr, 1963).

Internationally, a wide range of forms of assessment are used to assess practical skills. Abrahams, Reiss, and Sharpe (2013) distinguish between the *Direct Assessment of Practical Work*, which includes a practical exam, where students are observed carrying out a practical activity, and *Indirect Assessment of Practical Work*, where students may be assessed on the product of a practical activity (e.g., a laboratory report), or may be asked written questions in an exam. Although written questions about practical work cannot assess students' ability to manipulate equipment, they are frequently used to assess other aspects of practical work, including knowledge of experimental procedures and techniques, data analysis and presentation, and the interpretation of data with respect to scientific concepts. A secondary aim of the inclusion of written questions about practical work is to encourage the teaching and learning of practical work, because this will serve as a good preparation for the assessment. This will support the development of those skills (e.g., manipulative skills) which are not directly assessed. Written questions may be used as part of a written exam, or comprise a whole exam paper. Given the diversity of practical science skills which might be assessed, it is important to understand which skills are assessed and in what proportion in any given assessment. Teachers and students need to know how to structure teaching and learning to ensure that they are prepared for the assessment. Similarly, assessment organisations (AOs) that set the assessments must ensure that the balance of skills assessed is appropriate for the intended curriculum, both within a single assessment, and across the lifetime of a qualification.

Although individual syllabuses and curricula state the skills and knowledge which should be assessed in a specific qualification (e.g., the assessment of how observations are recorded, measured and estimated), currently there is no framework for categorising the skills and knowledge assessed using written questions about practical work which has been designed to be used to compare different qualifications, and used in different contexts.

This article describes a study in which the development of a taxonomy is first described, then its application in evaluating current science qualifications explored. The taxonomy aims to classify practical skills

assessed using written questions about practical work. The taxonomy seeks to provide an accessible description of the skills, knowledge, and understanding (constructs) which underlie practical science in written exam papers. A taxonomy of practical science skills on questions about practical science has the potential to allow evaluation and monitoring of practical science questions for AOs as it allows for comparisons of skills assessed over time, between papers and between subjects. Cambridge Assessment International Education assessments include written practical science questions (as an alternative to practical exams) and so an evaluation as to how the skills assessed have changed over time, and how they vary between subjects, would provide additional information on the performance of the assessments. OCR assessments of written practical science were introduced in 2016 as a result of reforms in the qualifications. Therefore, evaluating whether the assessments are similar to Sample Assessment Materials (SAMs), and how subjects compare to one another in terms of the skills assessed, would aid the meeting of regulatory requirements to evaluate and monitor the new assessments.

The study described here was undertaken in two phases:

Phase 1: Development of the taxonomy aimed first to develop a taxonomy, then to elicit feedback from stakeholders in order to determine its efficacy. It, therefore, entailed two stages:

1. Teacher and stakeholder input; and
2. Refining the taxonomy.

Phase 2: Application of the taxonomy had three aims:

1. To analyse the skills assessed in different papers, for different sciences and across several years;
2. To use data on the functioning of individual questions to evaluate the quality of questions assessing particular skills; and
3. To determine in what other contexts and for what benefits the taxonomy could be applied.

Phase 1: Development of the taxonomy

The focus of this study was General Certificate of Secondary Education (GCSE), General Certificate of Education (GCE) Advanced Level (A Level) and Advanced Subsidiary Level (AS Level) science question papers from one UK AO, and an international GCSE (IGCSE) written alternative to practical science question papers.

For the first phase of research, the papers analysed included:

- IGCSE *Alternative to Practical* (ATP) written question papers from June 2014 and 2015 for Biology, Chemistry and Physics. The ATP papers include experimental contexts covered by the practical assessment of the IGCSE science assessments, but are indirect assessment of practical skills; and
- GCSE and GCE SAMs for each of the three science subjects (designed to reflect recent reforms in the assessment of practical science). The questions in these papers are referred to as *embedded IAPS* (Indirect Assessment of Practical Skills).

Stage 1: Teacher and stakeholder input

Teachers were asked to categorise practical questions from science papers. Teacher-centred input was considered to be crucial, given concerns about the impact of assessing science practical work through written questions on classroom practical work. A group of nine science teachers was recruited (three Biology, three Chemistry and three Physics). Teachers had differing experience of teaching GCE AS Level, GCSE, and IGCSE.

Stage 1 comprised two sessions: a *subject group* session, and a *cross-subject group* session.

Subject-group session

Teachers were required to work in groups of threes, in subject specialisms (e.g., three Biology specialists working together). Each group was provided with packs of question papers appropriate to their specialism, with questions about practical work highlighted. Teachers were asked to think about how they would categorise the highlighted questions, in terms of the type of knowledge or skill that was being assessed, and to comment on any aspects relating to progression from GCSE and IGCSE to AS Level. Having decided on a list of categories, each group was then required to generate a short description of the category, such that another science teacher could use their categories to identify questions which fell into each category. Teachers were also asked to list any questions in their packs of papers that fell into each category. Finally, teachers were encouraged to log any issues or challenges arising whilst they worked, especially in relation to questions that were difficult to categorise.

Cross-subject group session

In a subsequent session, teachers were asked to work in groups of three, with one subject specialist from each subject comprising each group. This time, groups were asked to review the three category lists developed in the first session, and identify areas of convergence and divergence. Additionally, each group was encouraged to consider the questions identified for each category in the first session. The cross-subject session culminated in a plenary attempt to develop one set of generalised categories which could be used for all three science subjects. The set of categories helped inform the construction of a draft taxonomy of skills from written questions about practical work.

The workshop held with teachers (Stage 1) produced lists containing a total of 57 words and phrases; the Biology subject group produced a list of 23 words and phrases, Chemistry 14 and Physics 20 with very limited direct repetition. Separating individual actions from phrases which combined several skills allowed the creation of a combined set of 48 categories, which was subsequently distilled into 15 statements.

Stage 2: Refining the taxonomy

Stage 2 consisted of eliciting feedback on the distilled taxonomy (constructed in Stage 1) from stakeholder groups (including representatives from HE, subject associations, and teachers). These groups identified the need for segregation of drawing skills into two separate categories. Firstly, the accurate representation (and labelling) of objects observed, such as required in Biology or Geology and, secondly, the more abstract diagrammatic representation of objects using defined symbols, such as electrical circuits, molecular structures, or laboratory apparatus.

The distilled taxonomy developed in Stage 1 was also scrutinised by assessment specialists in each subject to categorise assessment items for a subsequent qualitative analysis. In carrying out this activity, any difficulties identified in assigning a category to a question or item were catalogued.

Table 1: Description of the taxonomy of practical science skills

Skill	*Brief explanation*	*Detailed explanation*
Diagrams (apparatus and circuits)	Representation of equipment or circuits using accepted symbols.	The ability to represent circuits or apparatus in accepted forms.
Drawing (Biology) and labelling	Biological drawing, accurate representation.	The ability to accurately represent objects observed and to appropriately label them.
Making measurements	Reading a scale from a diagram (prevalent in alternative to practical papers).	A practical test of the ability to use a scale to take a reading. In more complex scenarios, a time-lapse photograph or scale diagram may be reproduced requiring students to make measurements.
Recall	Application of knowledge.	A response to a question relating to a practical activity defined in the specification, drawing on recall of theory, or carrying out the activity.
Capturing data	Observing and reading data, interpreting data from a table or graph, recording data.	This is distinct from making measurements. In this case, the data is taken from a table or graph. As opposed to plotting a graph, in this instance the student retrieves data from the graph. The skill is also deemed to involve the ability to record the data appropriately.
Data analysis	Problem solving.	Having identified the appropriate data from a range of sources or different calculations, linking the information to allow the solution of a problem.
Data handling	Calculation (e.g., calculating gradient).	The use of data to carry out a calculation using a formula which may require rearrangement, or the linking of formulae.
Data interpretation/ Identifying trends	Plotting and interpreting graphs.	Transferring data from tabular to graphical format. Being able to identify trends from the graph (telling the story of the graph).
Data quality	Evaluation of data and conclusions.	Comparing the outcome of an activity with the anticipated or accepted outcome. Using numeric processes to comment on the quality of the data, with possible reference to the uncertainty of the process or measurements.
Experimental design	Method, planning and procedure including identification of variables.	An understanding of the processes involved allowing the identification of variables and the ability to propose an experiment to demonstrate the required hypothesis or outcome.
Predicting outcomes	Understanding processes.	Using knowledge and understanding of a process to anticipate the likely outcome of a given sequence of events. Often examined by giving a scenario for a practical activity, specifying a change in the circumstances, and asking for identification of any changes in the outcome.
Use of apparatus and techniques	Application of knowledge of practical skills.	Typical questions could be the evaluation of the use of specific equipment or suggestion as to possible improvements. This has implied understanding rather than straight recall.

Overall feedback from stakeholders, as well as comments on the taxonomy collected during a *Science Forum* organised by a UK AO, contributed to the refinement of the statements and explanations to produce the final taxonomy. Assessment specialists/forum participants contributed to a number of adjustments to the taxonomy. This was particularly important as the taxonomy was different in nature to earlier categorisations of practical work depicted in the literature: It differed against such headings as planning, observing, analysing and evaluating – labels which describe the nature of a question, rather than the skills which may be used in its completion. This review of the taxonomy, with a range of stakeholders, teachers, university lecturers, and representatives of the scientific community (via the Science Forum), also gave feedback leading to the addition of the abstract diagrammatic representation of electrical circuits and stylised two dimensional diagrams of chemical apparatus as a second separate drawing skill.

The final taxonomy is shown as Table 1.

Three key issues were raised by the activities to this point:

1. The segregation of activities related to data into four distinct categories;
2. The distinction between drawing an accurate representation and the use of symbols; and
3. The nature of the taxonomy.

Teachers had clearly identified a range of different skills pertinent to the use of data. Consequently, we were able to identify questions which linked directly to these skills. Examples of the categories can be demonstrated in the following questions:

Example 1: Capturing data/Data handling

This 4-mark item (shown in Figure 1) requires (i) that the candidate captures the appropriate data from the graph, and (ii) that they handle the data to calculate the gradient and hence determine the Young modulus of the metal.

23 The extension of a metal wire is x when the tension in the wire is F. The table in **Fig. 23.1** shows the results from an experiment, including the stress and the strain values.

F / N	x / 10^{-3} m	stress / 10^{7} Pa	strain / 10^{-3}
1.9	0.4	1.73	0.20
4.0	0.8	3.50	0.40
5.9	1.2	5.21	0.60
8.0		7.00	0.80
9.0	1.8	7.95	0.90

Fig. 23.1

(a) Complete the table by determining the extension when the tension is 8.0 N. **[1]**

(b) **Fig. 23.2** shows a graph of stress against strain for the metal.

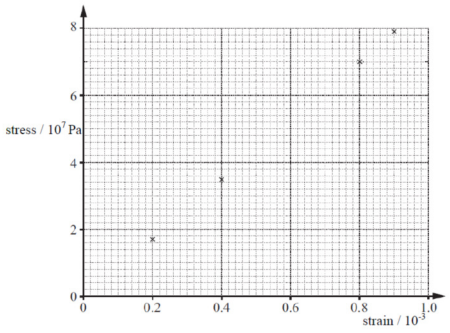


Fig. 23.2

(i) On **Fig. 23.2**, plot the data point corresponding to the tension of 5.9 N and draw the line of best fit through all the data points. **[1]**

(ii) Use **Fig. 23.2** to determine the Young modulus of the metal.

Young modulus = ...Pa **[2]**

Figure 1: Capturing data/Data handling

Example 2: Data interpretation/Identifying trends

The question in the next example (shown as Figure 2) is typical of those looking to assess the understanding of a practical activity by requiring the candidate to identify the trend of a graph to match the defined scenario. The other main question type in the data interpretation category is in transferring tabular data to graphical form, as incorporated in section (b)(i) of Example 1 (shown in Figure 1).

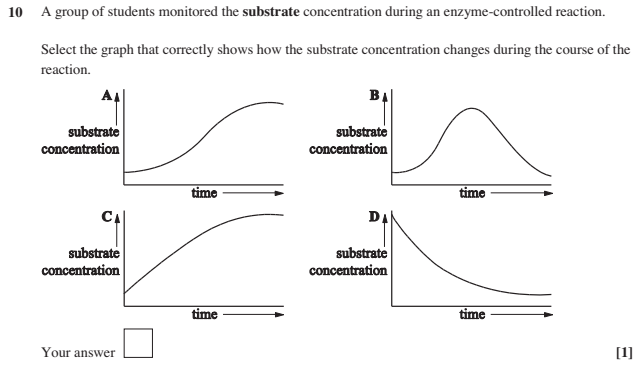


Figure 2: Data interpretation/Identifying trends

Example 3: Data quality

This is an open-ended question (shown in Figure 3) with a total of 6 marks available. It incorporates a range of lower order items from the taxonomy, which then contribute to the higher order. The question provides significant data and uses earlier questions relating to the data and graphs provided to subsequently contribute to the higher demand in making a comparison between the two possible methods.

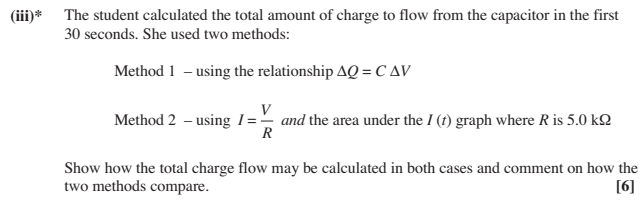


Figure 3: Data quality

Example 4: Drawing (Biology)

The review of an early iteration of the taxonomy identified the skills relating to drawing, evaluation and explanation as missing items. Quillin and Thomas (2015) discuss how drawings vary in the extent to which they are representational or abstract, both within Biology and the wider sciences. The following question (shown in Figure 4) depicts a biological drawing as an accurate representation of an object.

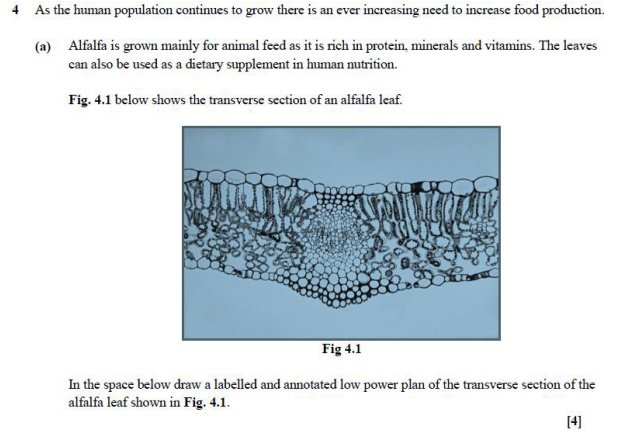


Figure 4: Drawing (Biology)

Phase 2: Application of the taxonomy

For the second phase of research, a separate set of papers (those taken by students in 2016) was used. This set of papers was the first assessment of the redeveloped international syllabuses, though there were no changes to the assessment of practical skills. Thus, the addition of the 2016 question papers allowed analysis of current UK and international assessment models.

The application phase sought to investigate whether the taxonomy could be applied to items on written practical science within past question papers and SAMs. The analysis, using the taxonomy, aimed to demonstrate how the taxonomy could be used to evaluate and monitor science practical questions by addressing two research questions:

1. *Are the skills in the taxonomy assessed?*

   This question was addressed by comparing the coverage of skills assessed in IGCSE, GCSE, and SAM items on question papers across years and subjects. This use of the taxonomy allows for AOs to monitor and evaluate whether assessments are consistent in assessing practical science skills.

2. *How do questions addressing particular skills perform?*

   This question was addressed by comparing whether particular skills from the taxonomy are associated with particular characteristics of item performance. This use of the taxonomy allows for AOs to evaluate whether particular skills have different characteristics in terms of difficulty and suitability of the item within the rest of the question paper. For instance, if certain skills have higher difficulty values than others, AOs may use this information to train question setters to recognise how difficult particular skills are, or to determine how skills should be assessed in the future. These analyses also provide evidence that the skills being assessed are appropriate to the assessment, in that they do not lead to high omission rates, or have a higher than anticipated difficulty.

Research question 1: Are the skills in the taxonomy assessed?

In order to address this question, subject specialists from both OCR and Cambridge Assessment International Education reviewed the question papers listed below. They first reviewed each item and judged whether it assessed a practical science skill, and if so, they then identified which skills were assessed using the taxonomy. Multiple skills could be assigned for each item, and the first skill assigned was used as the primary skill that the item assessed.

- IGCSE *Alternative to Practical* (ATP) written question papers from June 2014 – 2016 for Biology and Chemistry, and June 2013, 2015 and 2016 for Physics. These are referred to as ATP questions.
- GCE AS Level written question papers for June 2016 for each of the three science subjects (both specifications A and B for Biology and Chemistry, and Specification B for Physics). In addition, SAMs for Biology and Chemistry were also used. These are referred to as *embedded* IAPS questions.

As the subject specialists in this phase applied the taxonomy to items, they found them to be assessing practical science skills in past question

Figure 5: Practical skills assessed across the three sciences

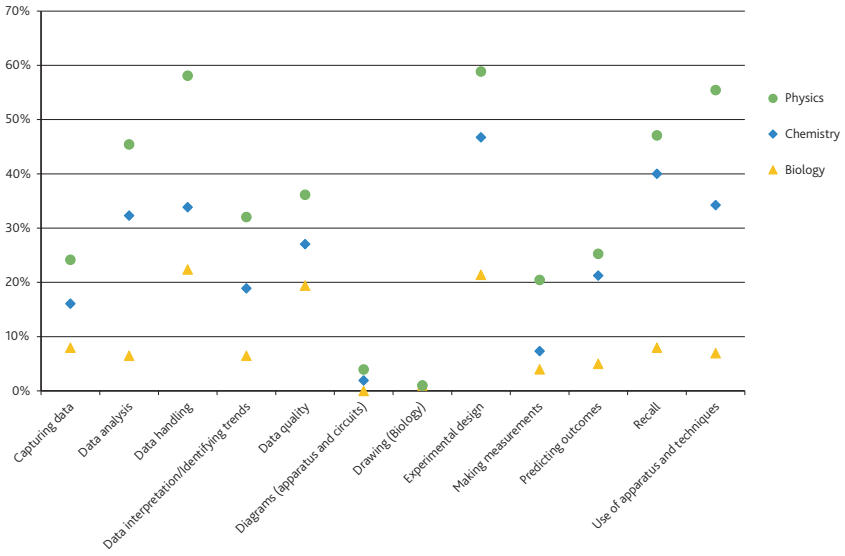
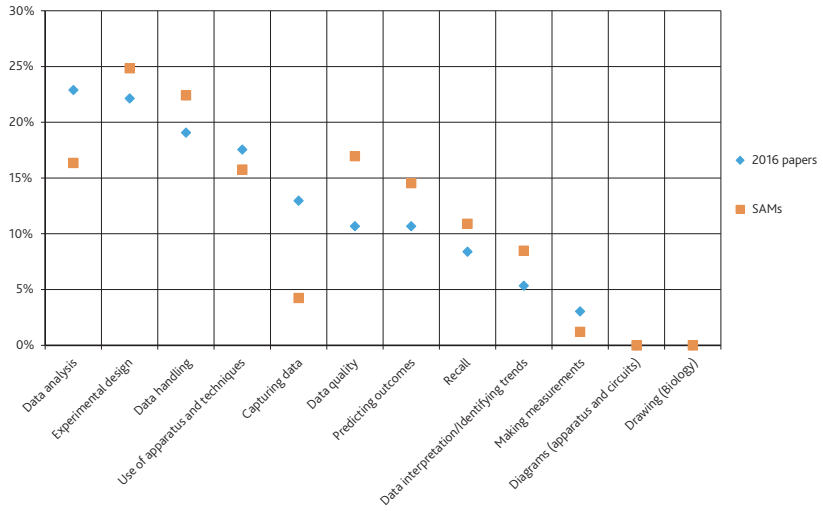


Figure 6: Practical skills assessed across the three sciences (SAMs and live papers)



papers, and analyses were carried out on the items identified. Firstly, the number of each practical skill assessed from all question papers was counted and comparisons made across years, AOs, and subjects. In addition to the planned comparisons, additional comparisons were explored in terms of comparing the mark tariffs of items assessing different skills, as a form of additional information that could be used to evaluate the performance of the assessments. Secondly, item level data was collected for all question papers investigated, and used in conjunction with the skills assigned to each item to compare performance measures of items according to the different skills they assessed.

All of the skills in the taxonomy were judged to be assessed to some extent on each paper analysed, although to varying degrees (see Figure 5). The most commonly assessed skills were *Experimental design, Data handling, Use of apparatus and techniques* and *Recall*. The popularity of Recall is probably impacted by the fact that half of the questions that assess this skill are also assessing other skills. The assessment of the very specific skills of *Diagrams (apparatus and circuits)* and *Drawing (Biology)* were rare.

In Figure 5, we show a comparison of the skills assessed by the three science subjects Physics, Chemistry, and Biology.

The proportion in which skills are assessed in the SAMs and the live 2016 papers are compared in Figure 6. This allows comparison of the 2016 live assessment against design criteria exemplified in the SAMs.

Proportions are very similar, with some discrepancies seen in *Data analysis, Capturing data* (fewer in SAMs), and *Data quality* (higher percentage in SAMs).

Research question 2: How do questions addressing particular skills perform?

To investigate how questions addressing particular skills perform, we analysed item performance using the three measures described in Table 2 for each skill and compared them to each other. As multiple skills were often assigned to an item, we used the primary skill assigned for

Table 2: Measures of item performance

Measure	*Description*
Facility	The mean mark on the item divided by the maximum mark was used to measure the difficulty of an item.
Omit	The proportion of students that did not attempt the item.
R_rest	The correlation between the item mark and the paper total excluding the item. R_rest is used to evidence whether items discriminate between good and weak candidates, and positive values indicate that pupils doing well on the question also do well overall.

each item. We used the facility, omit, and R_rest values calculated for all items that were assigned a practical skill from both ATP and embedded IAPS papers. SAMs were not used, as item level data is not available for these papers.

To compare how well items targetting each skill performed, the mean and standard deviation of facility, omit, and R_rest values of all items assigned to each skill was calculated. Figure 7 shows the mean performance of each skill, in terms of facility (Figure 7a), omit rate (Figure 7b), and R_rest (Figure 7c), demonstrating variation in facility and omit values depending on the skill assessed. The facility and omit values suggest two broad groups of skills that vary by their level of difficulty. Here, the skills *Making measurements*, *Drawing (Biology)*, *Capturing data*, *Data handling*, and *Use of apparatus and techniques* have the highest mean facility values and the lowest mean omit values of the skills, indicating that items assigned as assessing these skills are easier compared to other practical skills. In contrast, the skills *Data Analysis*, *Data interpretation/Identifying trends*, *Data quality*, *Experimental design* and *Predicting outcomes* have the lowest mean facility values, and the highest mean omit values of the skills, indicating that items assigned as assessing these skills are more difficult compared to other practical skills. Furthermore, the items assigned as assessing *Recall* and *Diagrams and circuits* do not fit well with either group based on their mean facility and omit values, and so are assigned to neither group. Firstly, for the skill Diagrams and circuits, this may be due to the high variation and a very small number of items assigned as assessing these skills. Next, for the skill Recall, this may be due to the broad range of questions in which this type of skill is likely to be assessed. Finally, there is little variation in the mean R_rest values of each skill, indicating that all skills are performing

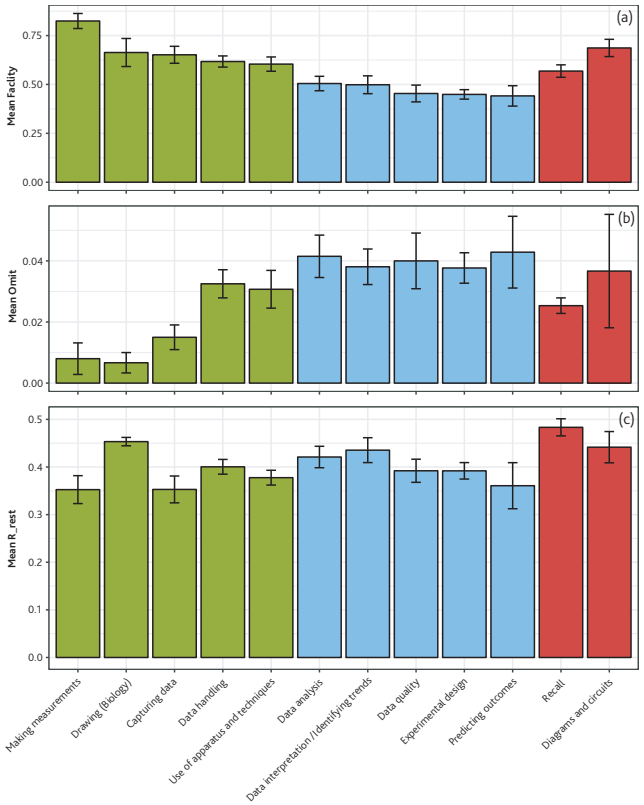


Figure 7: Measures of how well items assessing different skills functioned by subject, using both Cambridge International and OCR data
(Note: Mean facility values (a), mean omit values (b), and mean R_rest values (c), with standard error bars. The groups are distinguished by colour whereby the group with high facility and low omit are in green, the group with low facility and high omit are in blue, and the skills that do not fit in either group are in red).

to a similar level in terms of distinguishing between candidates. While the item performance measures varied between subjects (Biology, Chemistry, Physics), there was no consistent pattern whereby certain skills in certain subjects vary in comparison to other subjects or skills.

Discussion

The work described here constitutes an attempt to develop and apply a taxonomy of written questions about practical work. Locating practical science skills within the context of an explicit framework affords a more systematic and overall coherent approach to classifying and conceptualising such skills in written exam papers.

The integrity of a practical science test – irrespective of whether it attempts to assess practical skills *directly*, or whether such skills are *indirectly inferred* (as they might be in a written exam or through some other secondary form of assessment) – depends to a large degree upon a comprehensible understanding and articulation of the underlying construct(s) which it seeks to characterise. If these constructs are not well defined, then it will be difficult to support the claims a test developer may wish to make about the utility of the tests, including claims that the science tests do not suffer from factors such as *Construct underrepresentation and Construct irrelevance* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p.63).

A useful (and necessary) distinction should be made between what is easy to assess, and what is important to assess. It might be argued, for example, that data skills (*Capturing data, Data analysis, Data handling, and Data quality*) are easily assessed whilst also being considered crucial to the area of practical science work. Other skills subsumed under the classification *Conceptual understanding*, are often considered an important reason for doing practical work, though interestingly the literature on the efficacy of science practical work in augmenting the development of conceptual understanding is somewhat mixed (Hewson & Hewson, 1983; Hofstein & Lunetta, 1982; Lazarowitz & Tamir, 1994; Mulopo & Fowler, 1987).

Abrahams and Reiss (2015) state that indirect assessment of practical science is more appropriate for determining a student's understanding of a skill or progress, whereas direct assessment is more appropriate to determine a student's competency. The implication being that there is a potential danger that understanding of skills that are easy to assess are assessed frequently, at the expense of the understanding of other skills, and at the expense of competency. If this is the case, depending on the purpose of the assessment, an alternative form of assessment might need to be used in combination with the written exam.

How can the taxonomy be used, and by whom?

We believe that the taxonomy can be used in a variety of ways and by a range of educational practitioners (such as teachers, AOs, curriculum developers and test developers) to:

- provide a structure for classifying established, predetermined categories of indirect practical science skills that can be used by AOs and test developers for considering their intentions with respect to the assessment claims they wish to make;
- prompt an evaluation (on the part of the test developer) of how effectively the assessment claims have been met;

- allow test developers to construct question papers that elicit a range and balance of appropriate, effective practical science skills;
- ensure an appropriate level of predictability for those writing questions (both at question and paper level);
- monitor question papers over time;
- generate, from a regulatory perspective, SAMs representative of future live papers;
- afford an opportunity for test developers to consider their intentions with respect to comparability both *within* qualifications (e.g., across science subjects) and *between* qualifications (e.g., GCSE and its international counterpart);
- enable the efficacy of formative tasks to be determined (and evaluated) in relation to the purposes and objectives of the teacher; and
- offer test design and development practitioners a means for evaluating assessments from different (competitor) AOs.

Reflecting on how the taxonomy has been used

Since developing the taxonomy, it has been used in OCR to compare the coverage of assessment of skills across the three sciences on a routine basis. This has proved useful in that it has provided information for assessment specialists to help consider the consistency of skills coverage across the subjects. The taxonomy has also been used to compare skills coverage in the SAMs and live assessments. For example, the SAMs have recently been compared with the 2017 A Level. As the SAMS exemplify for schools the types of questions and coverage in the live assessment, it is very important that the live assessment reflects the coverage exemplified in the SAMs. In this context, the taxonomy has proved invaluable to OCR. More generally, AOs can use this information to support a good match between the SAMs and live materials, thereby supporting schools in their understanding of the expectation of the assessments.

The taxonomy has helped to establish a clear and well-articulated position on the underlying construct(s) of written practical science assessments. Having a clear understanding of how practical science constructs are conceptualised will serve to operationalise those constructs for assessment purposes in the future. As such, the taxonomy will support subsequent redevelopment and/or revision of science qualifications and provide a sound rationale for the proposed changes on construct and other grounds such as practicality, impact, validity and reliability.

Conclusion

The taxonomy provides a framework for considering which practical skills are assessed and how frequently they are assessed. However, it does not address the more fundamental question of which practical skills we should assess in a written paper, and what the relative frequency of assessment each skill should have. In this study, we have compared items used to assess two different levels (GCSE and GCE), and across Biology, Chemistry, and Physics. Should GCSE/IGCSE students be assessed on the same skills (and in the same proportions) as GCE students? The GCSE/IGCSE arguably serves a broader purpose (progression and development of scientific literacy) than the GCE, which is focused on preparation for HE. Should Biology, Chemistry, and Physics follow similar patterns, or do inter-disciplinary differences entail that different proportions are appropriate? It is beyond the scope of this article to answer these questions. However, by considering the taxonomy in relation to the purpose of a science practical assessment, we believe we have provided a tool with which to frame the debate.

Acknowledgements

We are indebted to a number of people who contributed to this study. We particularly thank Diana Marosi, formerly of Cambridge Assessment International Education, for her analytical input. We would also like to thank the assessment practitioners and teachers involved in the study for their input throughout the course of the work.

References

Abrahams, I., & Reiss, M. J. (2015). The assessment of practical skills. *School Science Review, 96*(357), 40–44.

Abrahams, I., & Reiss, M. J. (2012). Practical work: Its effectiveness in primary and secondary schools in England. *Journal of Research in Science Teaching, 49*(8), 1035–1055.

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.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

Hewson, M., & Hewson, P. (1983). Effect of instruction using student prior knowledge and conceptual change strategies on science learning. *Journal of Research in Science Teaching, 20*(8), 731–743.

Hofstein, A., & Lunetta, V. N. (1982). The role of the laboratory in science teaching: Neglected aspects of research. *Review of Educational Research, 52*(2), 201–218.

Kerr, J. F. (1963). *Practical work in school science*. Leicester: Leicester University Press.

Kirschner, P. A., & Meester, M. A. M. (1988). The laboratory in higher science education: Problems, premises and objectives. *Higher Education, 17*(1), 81–98.

Lazarowitz, R., & Tamir, P. (1994). Research on using laboratory instruction in science. In Gabel, D. L. (Ed.), *Handbook of research on science teaching and learning*. (94–128). New York: Macmillan.

Lunetta, V. N., Hofstein, A., & Clough, M. P. (2007). Learning and teaching in the school science laboratory: An analysis of research, theory and practice. In S. K. Abel & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 393–441). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Mulopo, M. M., & Fowler, H. S. (1987). Effects of traditional and discovery instructional approaches on learning outcomes for learners of different intellectual development: A study of chemistry students in Zambia. *Journal of Research in Science Teaching, 24*(3), 217–227.

Quillin, K., & Thomas, S. (2015). Drawing-to-Learn: A Framework for Using Drawings to Promote Model-Based Reasoning in Biology. *CBE Life Science Education. 14*(1):es2.

Wilson, F., Wade, N., & Evans, S. (2016). Impact of changes to practical assessment at GCSE and A-level: the start of a longitudinal study by OCR. *School Science Review, 98*(362), 119–128.

Data, data everywhere? Opportunities and challenges in a data-rich world[1]

Nicholas Raikes Assessment Research and Development

Introduction

Everyone reading this will have heard tell of both utopian and dystopian visions for how "big data" and machine learning will change our lives. We know of the stream of data we leave whenever we use our smartphones, of the vast oceans of data held by corporate titans like Facebook and Google. We have heard how this data is the "new oil": the fuel for ever more sophisticated artificial intelligences that will change the world.

You might have come across ALTSchool (http://altschool.com). It closed some schools and refocused its strategy in 2017, but an article in *Education Week* as far back as 2016 described "lab schools" where sensors, cameras and microphones captured every physical action, every social interaction of every child every day, to supplement data gathered as the children used learning software. The vision, according to Herold, was to create:

> *Data flowing from the classroom into the cloud … a single river of information. Data scientists would then search the waters for patterns in each student's engagement level, moods, use of classroom resources, social habits, language and vocabulary use, attention span, academic performance, and more.* (2016, para. 6)

Findings from the lab schools would steer the development of the learning platform and be applied in other schools.

Despite bold visions such as this, there has yet to be a big data revolution in education – but not every claim should be dismissed as hype.

Assessment today

Big international assessment organisations like Cambridge Assessment have long held considerable amounts of data. For a typical paper-based "high-stakes" assessment, such as the International General Certificate of Secondary Education (IGCSE) or General Certificate of Education Advanced Level (GCE A Level), we know background information about most candidates, such as their date of birth, gender and school; we have their detailed marks and grades on the assessments they take with us; we know the questions they answered and who marked them; and we have their handwritten answers (as scanned digital images) and multiple-choice test responses. We use this data, for example, to give detailed information to teachers on how their students performed on the different topics tested (Figure 1), and to provide detailed information to test writers on how their questions performed, so that they can write even better questions in the future (Figure 2). More recently, we have started to use machine learning in our quality control processes. For example, we have trained a model to identify markers who are likely to be stopped due to inaccurate marking, and deployed it to monitor marks returned online and "flag" potentially poor markers for early intervention.

In this way, we can spot and fix problems sooner than we otherwise would.

1. This is an edited transcript of a presentation given at the 2018 annual conference of the International Association for Educational Assessment, in Oxford, UK. It can be viewed as a Cambridge Assessment *Research Byte* at www.youtube.com/watch?v=8_FP6YDDJ1I&t=2s

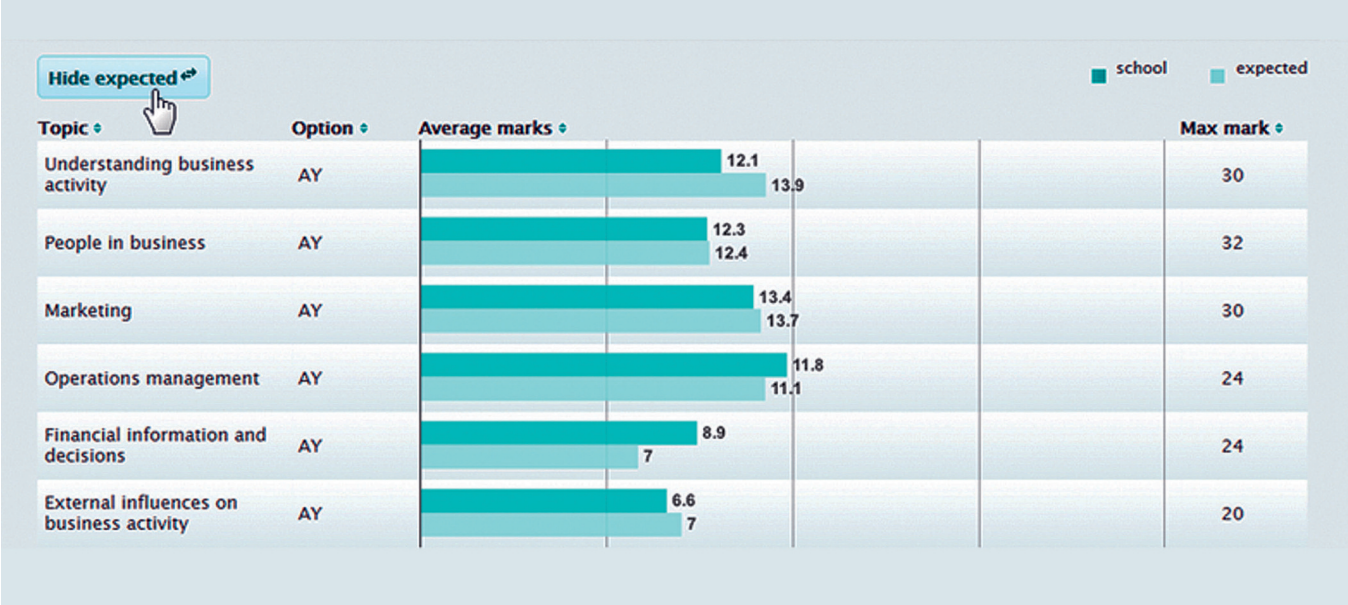


Figure 1: Results analysis for teachers
The dark bars show the average mark scored by a school's students on each topic within a qualification; the lighter bars show the average expected from a statistical model.

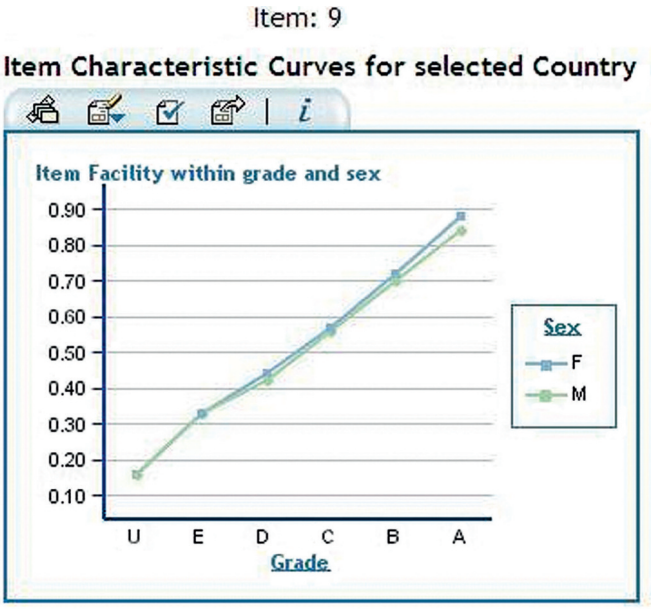


Figure 2: Question data for assessors
Detailed information is routinely provided to assessors to help with quality improvement. In this example, the chart shows how well male and female candidates with different grades overall performed on a given item.

We can even do an increasing amount with handwritten answers. Benton (2017) describes a method for spotting changes in handwriting between examinations, which could be due to an imposter sitting one (or a change of imposter). The two passages in Figure 3 were identified by the method – these were supposedly written by the same person, but this is questionable. In a personal communication to the author on September 4, 2018, Benton described how he had also demonstrated that it is possible to machine-read images of handwritten text with enough accuracy to detect some blatant cases of copying or collusion, though accuracy may depend on handwriting neatness. For example, the two passages shown in Figure 4 were detected automatically using Benton's method from amongst 18,000 others.

When text is produced digitally, we can do more with it. For example, we have operationalised computational text analytics in our question authoring and test construction processes. This allows us to screen draft exam questions automatically for any which are too similar to questions already published in text books. We also automatically screen reading passages for topic similarity in an automatically constructed, computer-delivered reading test, thereby ensuring that every student gets a variety of texts to read.

Surprisingly to many, there have been examples of automatic scoring of extended writing for around 20 years, though what works well in one context may not be applicable in all others. High-stakes tests of writing usually restrict automatic marking to providing a "second opinion" for comparison with human markers. The Cambridge Assessment English *Linguaskill* online test is used by organisations to check the English levels of individuals and groups of students, and contains a writing assessment which is automatically marked by "a series of computer algorithms that has learned how to mark test responses from a large collection of learner responses marked by expert human markers." (Cheung, Xu & Lim, 2017, p.3).

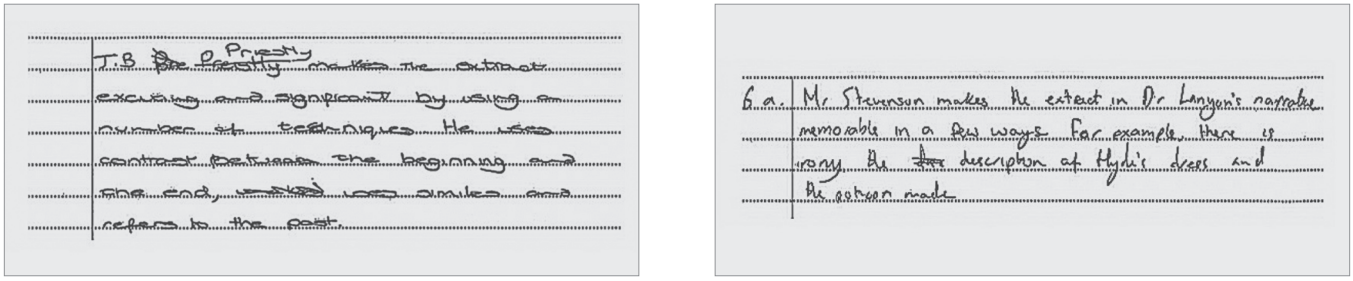


Figure 3: Malpractice detection in handwritten answers (Benton, 2017).
Both extracts were supposedly written by the same person, in different exams, but it is plausible that they are the work of different individuals. The change in handwriting was spotted automatically from changes in the median pixel density per word. Reproduced courtesy of the author.

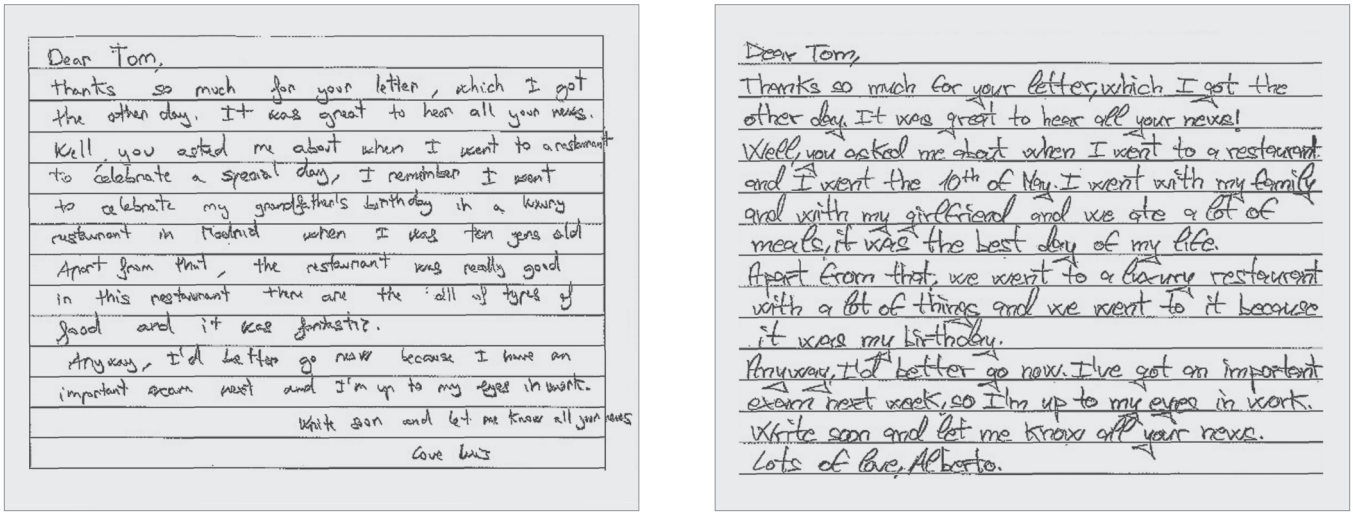


Figure 4: Malpractice detection in handwritten answers (T. C. Benton, personal communication, September 4, 2018).
Handwritten scripts were machine-read with enough accuracy to detect copying or collusion from amongst 18,000 other scripts. Reproduced courtesy of the author.

Opportunities

Let us turn now to the opportunities brought about by big data. There is no well-defined dividing line between big data and ordinary data, but big data is often considered to have three characteristics. In addition to *volume*, it has *variety* – encompassing text, video, images, log files of, for example, the key strokes and mouse clicks which students make as they engage with a computer-based test, and of the time spent focused on each task – as well as structured data, like marks and grades. This data might be streamed for analysis in almost real time at high *velocity*, which is the third characteristic of big data.

The technology and software for working with data are developing fast. Open source software such as Hadoop, Spark, R, and Python incorporate the latest advances almost as soon as they are made. Immensely powerful computing platforms can be built from relatively cheap hardware, or provided by cloud services such as Microsoft Azure and Amazon Web Services (AWS).

Machine learning has had some notable successes recently. Essentially, machine learning is statistical modelling rebranded and applied to automation. Arguably, some recent advances have had more to do with the increasing availability of data and processing power than with fundamental advances in the field of artificial intelligence. Nevertheless, these advances have potentially wide application.

In the remainder of this article, I will outline a couple of applications of big data in education, as illustrative examples of what might be possible, and finish by discussing some challenges to be overcome.

Formative assessment

The first potential application we will consider is in formative assessment; that is, assessment designed to guide learning. We would like to increase the amount of formative assessment which takes place, but high-quality questions are expensive to produce conventionally and, therefore, are scarce. However, teachers often write their own questions for use with their students. What if we provided teachers with an online platform on which they could upload their tests, and test their students, but which also made it easy for them to share and use each other's questions (items) and tests? As data accumulated, automated analytics could continuously refine estimates of item difficulty, and of how scores on one item related to scores on other items, opening the possibility that a machine-learning algorithm could be trained to categorise items, suppress bad ones, and help teachers select items and construct tests to meet their – and their students' – particular requirements. Moreover, if some of the item types were marked at first by teachers, the platform would accumulate data which could be used to train a machine-learning algorithm to mark the items automatically, thereby increasing still further the usefulness of the item bank to teachers and students.

Improved learning materials and personalised recommendations

As we collect data from more frequent testing, we will accumulate rich longitudinal data, reflecting each student's learning trajectory. We may have their work and detailed logs of what they did, as well as their marks. This would be a powerful resource for understanding learning progression and dependencies, and could be used to improve learning materials and develop better advice for teachers and students. If we pool data from learning management systems and from formative and summative assessments, we will be able to develop more intelligent adaptive learning systems. By combining the detailed information held on a student with machine-learning algorithms trained on historical data accumulated from many students, it might be possible to provide personalised recommendations automatically such as "Nick is relatively weak at algebra. This resource proved effective at raising scores for similar students".

Challenges

Most readers will by now be wondering about data protection and privacy. These are very 'hot' topics, partly because of the General Data Protection Regulation (GDPR) which came in to force across Europe in May 2018. Partly, also, because of the furore surrounding the London-based company Cambridge Analytica, and Facebook (see, for example, *The Guardian*, 2019), but also, I believe, because many people increasingly want more control over the data they produce, and of how it is used, and wish to feel confident that it is stored securely and will not be abused. Concerns are particularly acute for data about children. The case of *InBloom* is instructive (Singer, 2014). This was a project in the US with $100 million funding to create a detailed repository of educational data which would enable the kinds of application discussed above. It failed because of concerns, and then campaigns, about privacy and data protection, which snowballed until school districts and then states withdrew and the whole project collapsed. The failure highlighted the importance of establishing trust when undertaking projects such as this. This trust will depend, I believe, on having clear ethical principles and scrutiny; on being open; on communicating often and effectively to assuage concerns and inspire data subjects with the vision behind the work, encouraging them to see their data contribution as positively as volunteering or making financial donations to charity; and on gaining informed consent.

A less obvious challenge is statistical naivety. Limitations and caveats over statistical findings apply to all kinds of data, including big data. Correlation does not necessarily imply causation; data might not be representative of all people of interest, particularly if it is opportunistic; and important factors might be ignored because no data on them is easily available. Machine learning is applied statistical modelling. It is important to heed the wisdom of experienced statisticians and data scientists.

Machine learning is often described as a "black box". Models can be complex and involve many variables and levels of interaction. They are usually under the control of a data scientist, but can be hard to interpret in everyday terms. Algorithms may learn to be biased if trained on data which is biased. Microsoft's Tay chatbot was a famous early example of an algorithm skewed by training data, in that case due to pranksters feeding it extreme content on Twitter – see, for example, Lee (2016). Biases are not always so obvious, however. Sometimes algorithms are kept in black boxes for commercial reasons, their owners unwilling to be transparent about how they work.

For educational applications, we should insist on as much transparency as possible. For example, formative assessment is often described as "low stakes", but if a machine-learning algorithm gives poor advice to students, clearly their learning might be damaged, and the effects could be widespread if the algorithm is widely used. Students, like all human beings, do not always follow neat learning

progressions, making it hard to tell, for example, whether a poor performance in a short, formative test is a random aberration, or evidence of a fundamental misunderstanding and a real learning need. Developing tools of analysis and communication that can deal with this inevitable ambiguity is tricky. We should investigate the validity of machine-learning outputs, and whether they are aligned with alternative sources of evidence. And, we must evaluate the impact of data-fuelled approaches and machine learning products as they are introduced – and look for unintended consequences.

Cambridge Assessment has long been data driven. Big data, the convergence of teaching, learning and assessment, and the increasingly sophisticated operationalisation of machine learning and of data science more generally, are creating real opportunities for improving our understanding and practice of education. We should never put our faith in black boxes, however, nor introduce wide-scale change without evaluation. We must earn public trust by establishing and upholding clear ethical principles in relation to our use of data; be open; communicate continuously about what we are doing and why; inspire people with our vision and respond to their concerns; and always remember that we rely on their consent.

References

Benton, T. (2017). The clue in the dot in the 'i': Experiments in quick methods for verifying identity via handwriting. *Research Matters: A Cambridge Assessment publication, 23*, 10–16.

Cheung, K., Xu, J., & Lim, G. (2017). *Linguaskill: Writing Trial report*. Retrieved from https://www.cambridgeenglish.org/Images/466042-linguaskill-writing-trial-report.pdf

The Guardian (2019). *The Cambridge Analytica Files*. Retrieved February 8, 2019 from https://www.theguardian.com/news/series/cambridge-analytica-files

Herold, B. (2016, January 11). The Future of Big Data and Analytics in K-12 Education. *Education Week, 35*(17). Retrieved from https://www.edweek.org/ew/articles/2016/01/13/the-future-of-big-data-and-analytics.html

Lee, D. (2016, March 26). Tay: Microsoft issues apology over racist chatbot fiasco. *BBC News*. Retrieved from https://www.bbc.co.uk/news/technology-35902104

Singer, N. (2014, April 21). InBloom Student Data Repository to Close. *The New York Times*. Retrieved from https://bits.blogs.nytimes.com/2014/04/21/inbloom-student-data-repository-to-close/

Moderating artwork: Investigating judgements and cognitive processes

Lucy Chambers, Joanna Williamson Research Division and **Simon Child** Cambridge Assessment Network

(The study was completed when the third author was based in the Research Division at Cambridge Assessment)

Introduction

For the majority of standardised summative assessments in the UK, candidates will sit examinations. However, for certain practical or performance-based components, candidates will complete a non-exam assessment, which is marked by their teachers. To ensure that the standards of marking are the same across centres[1], samples of candidates' work from each centre are externally moderated. This process entails moderators, appointed and trained by awarding organisations, viewing the work and deciding whether the teachers have marked accurately and consistently. The aim of this study was to explore the cognitive processes and resources used by moderators when making judgements about artwork submitted for moderation.

The moderation method used by awarding organisations in the UK is that of inspection (see Joint Council for Qualifications[2], 2018, for a description of the moderation process). When making their judgements, moderators must consider the sample in the context of the centre as a whole, looking for trends and patterns in the marking. The moderators can make adjustments to the centre's marking, if necessary, to maintain the same marking standard across all centres. This must not be done with a view to changing the marks of individual candidates in isolation, but with a view to ensuring that the agreed standard is applied to all candidates (see Gill, 2015) for details of how centre-level mark adjustments are made).

Few studies have explicitly examined the cognitive processes involved in moderation. The only such studies that we are aware of are those of Crisp (2017) and Cuff (2017). The components under consideration in these studies involved the submission of mostly written work. The aim of this study was to investigate whether their findings hold when moderating submitted work of a very different nature, namely for Art and Design. There is little research on the marking and moderation of artwork. In fact, reviews observe that there is little detailed or technical research on assessment in art altogether (Gruber & Hobbs, 2002; Haanstra, Damen, Groenendijk, & van Boxtel, 2015; Herpin, Washington, & Li, 2011; Mason, Steers, Bedford, & McCabe, 2005).

Subject-specific research is particularly necessary for assessment in Art and Design. Assessment in Art and Design subjects is difficult: the skills involved in arts subjects are themselves complex, and furthermore "there exist many different conceptions of these skills" (Haanstra et al., 2015, p.413). Haanstra et al. go as far as to claim there is "no consensus on educational standards in the arts" (Haanstra et al., 2015, p.413). The particular demands of assessment in arts generally mean that the "forms and models of assessment particular to other areas of learning" do not transfer satisfactorily to Art and Design subjects

1. The vast majority of examination centres are schools or colleges.
2. The Joint Council for Qualifications (JCQ) is a membership organisation comprising the largest qualification providers in the UK. One of its aims is to provide common administrative arrangements for examinations.

(Eça, 2002, p.1). A consequence of this is that processes and concepts to do with evaluating assessment quality in other areas of learning also do not transfer directly to Art and Design subjects.

The Art and Design qualification used in this study contained two tasks, and could comprise of a variety of different art forms (e.g., annotated sketchbooks, mounted sheets, maquettes, prototypes, scale models, or written work). The candidates' non-exam assessment work (their submission/submitted work) were marked by the candidates' teachers. A sample, specified by the awarding organisation, of each centres' candidates' submissions, was then submitted for external moderation.

Previous moderation studies

Crisp (2017) used a *think aloud* method and moderator interviews to investigate moderation of General Certificate of Secondary Education (GCSE) assessments in English/English Literature, Geography, and Information and Communication Technology. Crisp described two groups of cognitive processes involved in moderation: (a) "reading and comprehending" the candidate work, and (b) making "evaluative judgments of quality" (Crisp, 2017, p.34). In terms of reading and comprehending, Crisp found that, in comparison with teachers, moderators were "more likely to make strategic choices about the level of detail in which they read different parts of students' submitted work" (p.34). Nevertheless, Crisp concluded that "the process of reading and understanding … appears to be unproblematic". This is in contrast to the evaluative processes, which are "rather more complex" (p.34).

The subprocesses that Crisp (2017) identified within the evaluative processes included attending to and evaluating features of candidates' work in relation to the marking criteria, indicating an analytic approach (p.34). However, Crisp also found that "most moderators appeared to apply 'configurational' processes in parallel", whereby "overall judgments are made directly and then checked against specific criteria" (pp.34–35). Reassuringly, Crisp found "no evidence of attention being paid to inappropriate features" of candidates' work and concluded that there was "no evidence of bias in judgments" (p.34). There was some evidence of affective reactions to candidates' work, but "these did not seem to influence judgments" (p.34).There was also evidence of comparative processes, these included comparison of a candidate's work to work by other candidates, comparison of candidates' work to other examples from the same candidate, and a tendency to arrange candidates' work in mark order.

The study by Cuff (2017) also used think aloud and interview methods. He used four specifications: GCSE History, GCSE English, GCSE Business Studies, and a Level 3 Extended Project Qualification. The aim of the research was to focus in greater depth on the cognitive processes involved in moderation, and on how moderators used possible supporting resources. Cuff noted that, in terms of the overall series of steps identified, "Encouragingly, these … align well with those reported by Crisp" (p.8). Many of the subprocesses that Cuff (2017) identified also aligned well with the details reported by Crisp (2017). In terms of resources, Cuff found that when reading the work, moderators formed impressions based on the marking criteria, previously moderated candidates, their understanding of the grade levels, and teachers' annotations written on the work. Crisp (2017) also found that moderators made use of annotations and the marking criteria when evaluating candidates' work.

Cuff found that "several aspects of the current findings suggest risks of confirmation biases in moderators' judgments" (p.35), which appears to contrast with the conclusions of Crisp (2017). The aspects that Cuff identified as potential sources of bias were an "anchor-and-adjustment" approach to adjusting marks (i.e., assuming the marks given by the centre were correct, unless shown otherwise), and the influence of moderators' initial impressions on their later judgements, even if the moderators themselves did not believe their judgements to have been affected.

Cuff (2017) recommended further research to confirm whether his findings applied "across a range of different contexts or where differences may exist (and why)" (p.37). To this end, we sought to explore the moderation process in terms of moderators' cognitive process and resources drawn on when making judgements about Art and Design submissions. Findings can contribute to the overarching moderation picture and help inform future training and moderation practice.

Method

The artwork used in the study was candidates' work submitted for an Art and Design qualification for 14–19 year olds. The Art and Design qualification contained two tasks: a Portfolio (worth 60%), and an externally specified Set Task (worth 40%). Both tasks were internally assessed by the centre and externally moderated. There were four Assessment Objectives (AOs) which were weighted evenly within each task. Five areas of study were available to candidates: Fine Art, Graphic Communication, Photography, Textile Design, and Three-Dimensional Design. Submitted work had to be in an appropriate format for the area of study and could take the form of, for example, annotated sketchbooks, mounted sheets, maquettes, prototypes, scale models or written work. The assessment of artwork for this qualification required holistic consideration of each candidate's submission, with marks assigned to each task.

The submitted work was sent to a central location where both the "live" moderation and this study took place. The researchers attended the standardisation meeting and observed some live moderation to enable us to mirror the live conditions as much possible. This study was conducted under experimental conditions several weeks after live moderation; this was because we did not want to disrupt live processes, nor risk affecting candidates' outcomes.

The study participants (N=3) were recruited from the small pool of moderators who had moderated the qualification in 2017. The participants all had significant teaching experience (15+ years) in Art and Design, and had taught Entry Level, GCSE, and General Certificate of Education Advanced Level (GCE A Level). They had all been or were currently Heads of Department and, at the time the study was conducted, all the participants held senior moderating positions.

Work from four centres was chosen, with each sample containing work from between two and eight candidates. The participants were instructed to moderate centre work in the same way that they would have done in the Summer 2017 session, using the evidence of candidates' work and resources that were available to them. Moderation was conducted for one centre at a time. The moderation task was to determine whether the specified marking criteria had been satisfactorily applied. In essence, this meant assessing whether the rank order of the centre sample was correct, and whether the marks given to the candidates' submissions were acceptable or would require adjustment. Participants were asked to record their marks and notes, as they would normally, and then write a report for the centre.

During moderation, a concurrent think aloud method was used.

The aim was to provide insights into the cognitive processes underpinning a specific activity through a verbalisation procedure (van Someren, Barnard, & Sandberg, 1994). The main advantage of this approach was that it provided researchers with additional information that would not be available through observation alone.

Prior to moderating, the participants were given a familiarisation task to give them the opportunity to get used to the think aloud method. The participants were provided with documents which replicated the materials that they would have had access to in live moderation. They comprised: a booklet of photographs that had previously been taken by the research team to represent the displayed artwork the moderation team had used during standardisation, a standards booklet (reference guide for moderators containing candidates' submissions benchmarked from across the mark range), a mark sheet, a copy of the marking criteria, a copy of the recording sheet on which moderators make notes of their observations, and a copy of the centre report template. The mark sheet contained the original marks for each candidate grouped by centre (total mark and mark by AO for each task).

The participants were asked to conduct moderating activities for approximately 90 minutes. They were allowed to take breaks at any time and it was made clear that they did not have to complete moderation for all centres. Participant activity was recorded via Morae software (TechSmith, 2011) and was observed by two members of the research team. The researchers sat beyond the participant's immediate line of sight. They recorded any relevant activities using observation schedules and noted anything that would comprise part of the interview to take place later in the day.

To account for the possibility that some of the participants might be more effective at verbalising their thinking than others, a retrospective interview was conducted with each participant after moderation had been completed. This was audio recorded. The aims were to illuminate and expand on think aloud outcomes, to add some information about the participants' thought processes, and to validate the researchers' early interpretations of the data collected.

Analysis

The recordings of the participants' spoken thoughts and activities from the moderation sessions were loaded into MAXQDA (VERBI Software, 2017). The research team familiarised themselves with the recordings by watching them and aligning them with the observation schedules coded during the observation period. An initial coding framework was developed with the aim of capturing the key activities; four categories of participant activity were identified:

1. Judgements about a candidate's level/mark;
2. Reference to documents;
3. Movements from submission to submission; and
4. Movements within a submission.

These broad categories were subdivided into several subthemes. For example, in the *Movements within a submission* category, the subthemes included leafing through work, leafing through work then focusing on one image, observation[3], speeding/leafing through a sketch book, consideration[4], lift up/bend down and touch/rotate work. The coding scheme was tested and refined. The researchers then double-coded (non-blind) all the data produced by the participants. This aimed to ensure consistency of application of the coding framework. Any disagreements between researchers were discussed and addressed. Typically, this took the form of a missed code. Within MAXQDA, it was possible to designate how long each coded activity lasted.

The interviews were first transcribed in MAXQDA, then analysed thematically.

Development of the process model

From the coded recordings of the moderation sessions, we developed a process model to describe how the participants carried out moderation. Firstly, for each separate moderation session, the codes described above were mapped against time. The timelines covered the period from the start of the moderation session (no work had yet been viewed and no other preparation work had yet begun), to the point at which the participant was ready to write the moderation report. Simplified timelines were also created (from the fully coded timelines) to show the candidate work each participant was viewing throughout moderation. For illustration, an annotated simplified timeline is shown in Figure 1.

3. Observation refers to comments on a candidate's/multiple candidates' use of styles, techniques and artists – all made without judgement
4. Consideration refers to comments with some element of judgement about the quality or realisation of the work of a candidate/multiple candidates. It denotes deeper engagement with the submission.

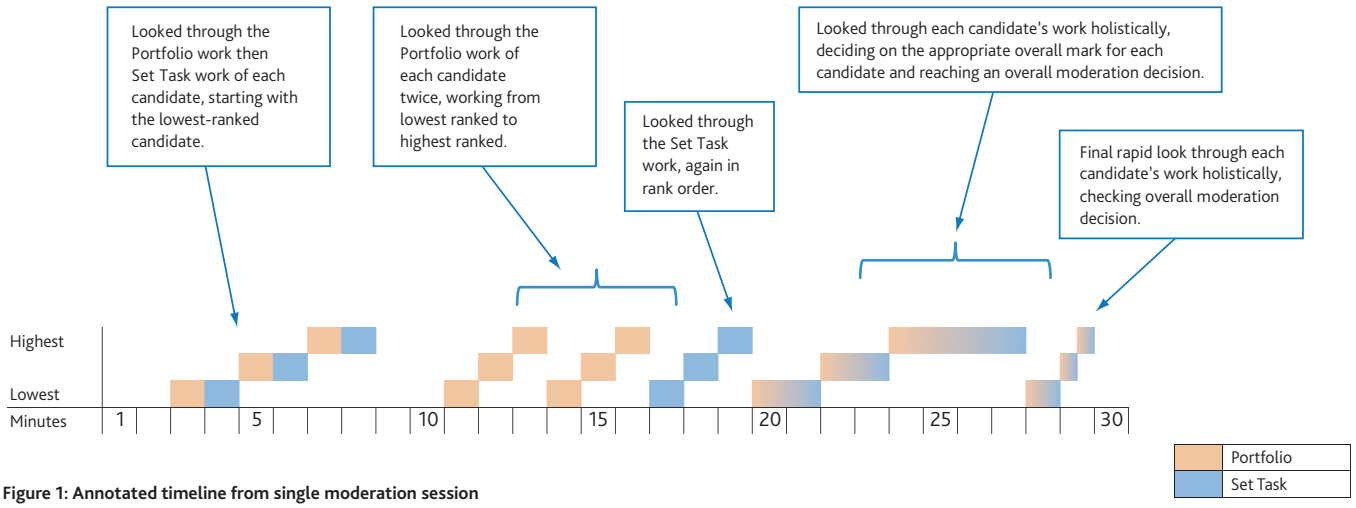


Figure 1: Annotated timeline from single moderation session

Secondly, the coded timelines were then compared to the moderation models of Crisp (2017) and Cuff (2017), to see how far these existing models were applicable. We found that while these existing models did not fully or accurately describe the observed moderation, certain features were evident. Although we could not use either model "as is", the models proved to be a useful frame on which to develop our artwork moderation model.

Thirdly, the coded timelines for each moderation session were synthesised into annotated diagrams, which were further combined into a single overall representation of moderation. The different stages of the process model were developed by: identifying repeated and systematic occurrences of distinctive combinations of cognitive processes, physical activities, and resources. For example, the familiarisation stage was characterised by rapidly looking through candidates' artwork throughout the sample, with high-level observations about themes, artists and techniques; some inferences about the course and/or centre; initial evaluation of the sample; and using the mark sheet to arrange work in mark order.

While there were some variations in the order in which the participants viewed work, the cognitive processes, physical activities, and use of resources formed coherent and identifiable stages that were common across all three participants – the differences in order did not necessitate separate process models for each participant.

Findings

Figure 2 shows the process model; we start by describing the model and then explore the differences in activities and cognitive processes that lie behind the moderation stages. The model is arranged in four columns: The first column shows the overall stages of moderation: orientation and preparation, familiarisation, investigation, reaching a moderation decision, and report writing. The second column shows the observed activities associated with each stage. The third column shows the cognitive processes associated with the stage (inferred from think aloud data), and the final column shows the resources drawn upon at each point.

Throughout the model, dotted lines indicate elements that varied among the three participants. For example, for two participants, the activity of setting up the moderation recording sheet occurred during the orientation and preparation stage, but one participant set up the recording sheet only after the familiarisation stage.

The next sections describe the stages of the model; inserted quotations illustrate activities that were typical in the different stages.

Orientation and preparation

The first stage of the observed moderation was an orientation and preparation stage. During this stage, the participants orientated themselves to high-level features of the centre and sample. In particular, from looking at the mark sheet, they noted how many candidates were in the sample, the marks given by the centre (centre marks), the rank order of candidates, and any unusual features. They also determined an order for the physical layout of candidates' work.

> *The candidates in this line up–and its centre XXXXX–there are three candidates. And … [writing down the candidate numbers] the marks are–the total marks are 64, 51, and 40.* (Participant C)

Two of the three participants prepared the recording sheet during the

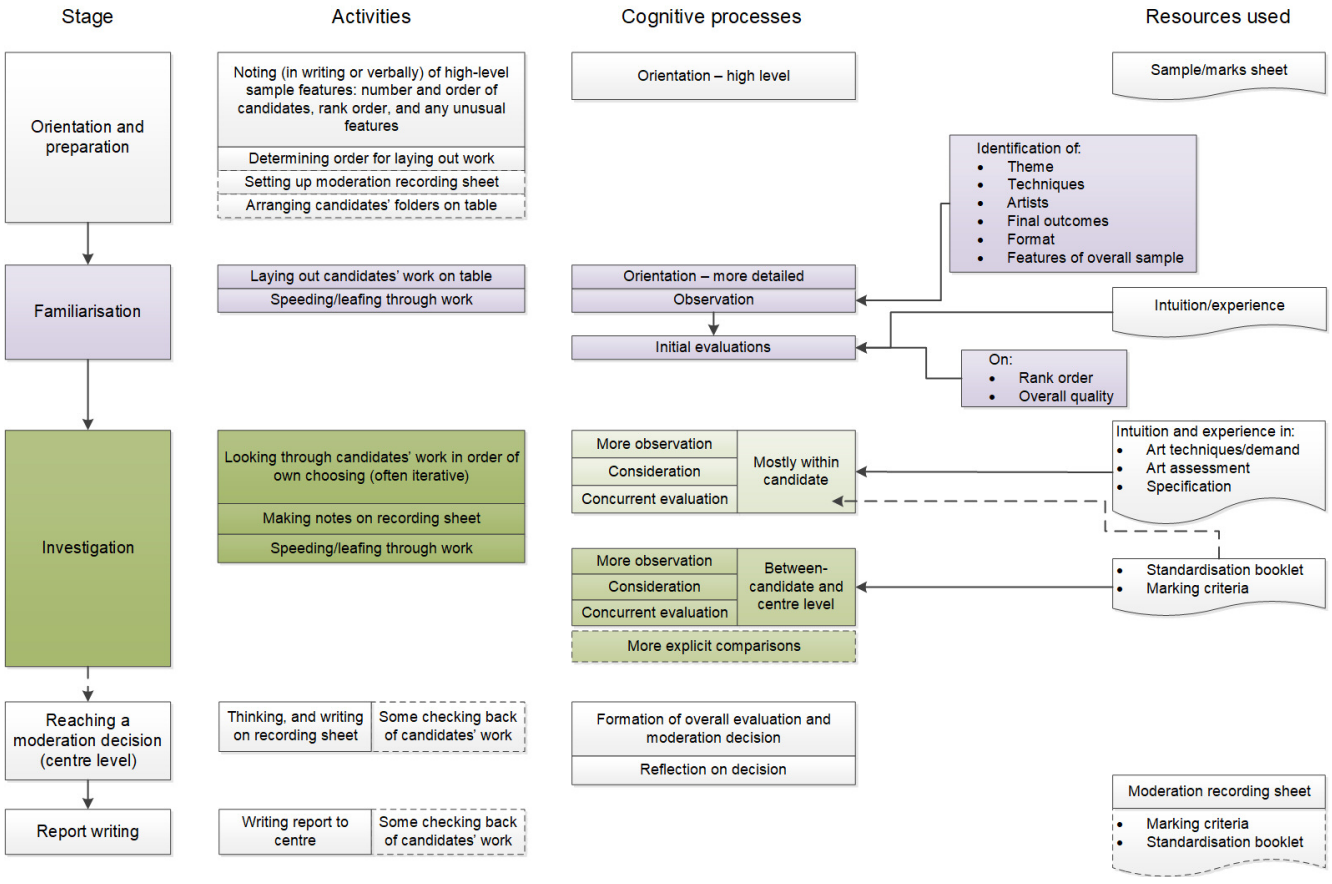


Figure 2: Artwork moderation process model

orientation and preparation stage. They added the centre details to the blank recording sheet, and transferred the marks from the mark sheet.

Familiarisation

The key activity carried out in the familiarisation stage was laying out candidates' submitted artwork on the table, in the order determined in the orientation and preparation stage. The artwork was spread out so that individual sheets of work were all visible, wherever possible. The participants each used a slightly different layout, but all were based on the rank order of candidates' centre marks. Each participant's chosen layout maintained a distinction between the Portfolio work and Set Task work of each candidate.

Whilst laying out the candidates' work, the participants leafed through candidates' work fairly rapidly. During this activity, the participants made observations on the theme(s) of the submitted work, the artists that the candidate showed evidence of studying, and on the techniques demonstrated by the candidate. The participants identified the format of the submitted work, and the "final piece" (final outcome) within the Portfolios and Set Tasks.

Besides observations about particular candidates or particular pieces of work, the participants made observations about features of the centre's sample overall. These observations included comments on the similarities and differences amongst candidates:

> *So, they've all done the same project …* (Participant B)

> *And what's interesting already by candidate … number two, I can see that–in terms of the procedure–of how the centre has got their candidates to produce their work, that the Portfolio is a study sheet with a sketchbook, and … the externally Set Task is … the externally Set Task is just using study sheets.* (Participant C)

In some cases, the participants needed to revise initial expectations and judgements as they progressed through the sample. In this quotation, the participant had (from the first two candidates' folders) concluded that all candidates in the sample had completed Portfolio work on the same theme, but revised this upon first looking at the next candidate's folder:

> *And, I bet this is space again! … Now, let's have a look at this [laying out candidate's work on the table]. Oh no! Mechanisms. So this candidate has worked–probably slightly more independently, chosen their own … I'm just supposing.* (Participant A)

Another distinctive cognitive activity that characterised the familiarisation stage was orientation to the submitted work. In contrast to the orientation activity within the *Orientation and preparation* stage (which could be summed up as orientation to *whom* the participants were moderating), the orientation within the familiarisation stage equipped the participants to know *what* they were moderating. The participants familiarised themselves with what had been submitted, and, crucially, how the individual materials within candidates' folders related (or did not relate) to the course requirements and marking criteria. Having briefly viewed the whole sample's Portfolio work, one participant explained:

> *At this stage, what I do is just make some preliminary notes, based on what I can see. So, we've got two distinct projects… and the first one is of 3D model making, which is about letters …* (Participant B)

In the following quotation, the participant explains how the laying out of the candidates' work has resulted in a physical overview, enabling location of work relating to specific AOs, at the same time as seeing this work in relation to the "final outcome" work:

> *I can see clearly now where the individual AOs are, and how they relate to each other. So I can see the 'Explore', just in this overview, and how it's impacting on the outcomes.* (Participant C)

The final cognitive activity that took place within the familiarisation stage was initial evaluation. Drawing on their intuition and experience (not yet on the formal resources of marking criteria or standardisation booklets), the participants made observations on the quality of submitted work. The participants themselves used the phrase 'first impressions', emphasising that these judgements were open to revision at a later point, and differed from the final professional judgements that the moderation process worked towards:

> *First impressions are that it looks strong.* (Participant B)

> *First impressions are that it looks terr–it looks under-marked …* (Participant A)

> *It's got a mark of 51–Level 1–and this is definitely higher than Level 1.* (Participant A)

> *Some good lettering there … This is definitely more than just into Level 2.* (Participant A)

During the familiarisation stage, the participants carried out an initial review of the rank order through laying out and looking at candidates' work in order of mark. None of the participants commented during the familiarisation stage that the centre's rank order was incorrect. However, if the participants' initial impressions of the submitted work were incompatible with the centre's rank order, it would necessarily have become apparent by the end of the familiarisation stage.

One participant described the activities of the familiarisation stage as indicating the "flavour" of what had been submitted for moderation:

> *By walking up and down the line I am actually registering the work, the standards of the work, I am actually getting that overview immediately and it means then, if you like, that when I come back to go through in terms of centre marks and the assessment, I've already got the flavour of what I'm looking at.* (Participant C)

This quotation conveys the participant's sense that the familiarisation stage provided the foundations for the later stages of moderation, during which detailed consideration of "centre marks and the assessment" would occur. A quotation from a different participant, at the end of laying out the whole sample's submitted work, similarly underlines the role of the familiarisation stage as a preparatory stage, and separate from the "actual" moderation process:

> *So, that's that. So, moderating. We'll get into moderating now.* (Participant A)

This distinction on the part of the participants supports the separation of the familiarisation and investigation stages in the process model. Although there was overlap in the types of activity observed within the familiarisation and investigation stages, the details and purposes of the activities observed were different, and the participants themselves appeared to consider them as separate stages.

Investigation

The primary activity of the investigation stage was looking through candidates' submitted work. The participants continued to make observations about the work, specifically, more extensive and more detailed observations on the same set of characteristics noted in the familiarisation stage. The participants also considered the quality of work, made evaluations of the work in comparison to the centre marks, and made explicit comparisons between examples of candidates' work, including those in the standardisation booklet. The participants made notes on the moderation recording sheet.

> *And this is a far more substantial portfolio of Set Work [Set Task]. He's looked at [Artist name 1], he's done mechanisms, and also … [Artist name 2].* (Participant A)

> *I would say that in terms of the sophistication, the confident experimentation, I would say … the quality of some of this drawing is very strong.* (Participant B)

> *Well actually, this set of work … doesn't quite have the achievement of the Portfolio. … And looking at our Set Work standards, and I'm firstly comparing–excuse me–to the 28 … It's slightly better than the 28, doesn't get to the 33 in our standards booklet.* (Participant C)

The first clear resource drawn upon during the investigation stage was participants' intuition and experience. Their knowledge and experience of art techniques (particularly the skill required to achieve particular outcomes), art assessment, and the course requirements were drawn upon frequently. The external resources drawn upon during the investigation stage were the marking criteria, and the standardisation booklet containing examples of candidates' work at particular levels, as referred to by Participant C above.

> *And that could go up. I'm going to–I'm just going to go and check my Level 2 criteria …* (Participant A)

> *Looking at the exemplar on 52, I would say this is stronger than the 52.* (Participant B)

The investigation stage was the most complex and lengthy of the moderation processes observed. The participants varied in terms of the order in which they looked through candidates' work, the precise point at which they drew on external resources, and the number of times they viewed the total sample. Participants viewed the work in rank order – one from the highest ranked candidate and two from the lowest. For two participants, the investigation stage was a highly iterative stage, and they each looked through the whole sample multiple times.

For all the participants, there was a gradual shift in the content of the considerations and evaluations of candidates' work that occurred during the investigation stage. As Figure 2 suggests, cognitive activity in the earlier stages of the investigation stage tended to focus on particular candidates. Towards the end of the investigation stage, cognitive activity more frequently focused on between-candidate comparison and centre-level consideration. Overall, a characteristic of the investigation stage was that the participants moved from detailed consideration of particular candidates' work, towards a point where they were ready to reach a moderation decision on the centre overall.

Reaching a moderation decision

The stage of reaching a moderation decision was the culmination of the investigation stage. The cognitive activity of this stage was forming an overall evaluation of the centre's marking, and reaching a decision on which marks to recommend.

The participants differed in the length of time they spent at this stage. For Participant C, the overall evaluation had been built up during the investigation stage – to the extent that reaching a moderation decision consisted of little more than stating and writing down the overall judgment. For Participants A and B, more time was taken, and there was a more sustained period of checking or reflecting on the decision. Both Participant A and Participant B looked through the entire sample again (Participant A did so twice) during the process of reaching and confirming their overall moderation decision.

> *The effect of the moderation so far is that we're moving some of the marks from where they were in Level 1 up to Level 2.* (Participant B)

Report writing

In the final moderation stage, the participants wrote a moderation report for the centre. Observation revealed that all the participants drew upon their moderation recording sheet to write this report, and some of the participants also referred to the marking criteria and standardisation booklet.

Aside from the activity of writing, the other activity observed during the report writing stage was some checking back to specific aspects of the submitted work. These checks were typically brief (sometimes just a glance) and often served to confirm a specific aspect of submitted work that the participant had referred to on the moderation recording sheet.

Discussion

Before we discuss the findings, it should be noted that there are some limitations to this research. First, it is necessary to exercise caution in generalising the findings given the small sample size. Although only three participants were involved, this was 60 per cent of the (small) population that moderated the target qualification. With only three participants, it is possible that individual differences could account for some of the findings. The study attempted to replicate live moderation as much as possible. Participants, however, were fully aware that this was a research exercise and their decisions would not contribute to candidates' results. In addition, the think aloud method used in the present research might potentially have increased the cognitive load for the participants, which might have influenced their moderation activities. However, it should be noted that when the participants were asked if the think aloud method disrupted their moderation activities, they perceived that it did not do so. This is in line with research by Greatorex and Suto (2008), in their study of the marking processes of 12 GCSE examiners. They found no relation between the type of items being marked (as a proxy for task difficulty) and the perceived ease with which the participants were able to think aloud.

To summarise our results, we found that the observed moderation process began with an orientation and preparation stage, followed by a familiarisation stage and then featured a lengthy investigation stage. This was followed by a stage in which a moderation decision was reached, and then finally a report writing stage. In terms of cognitive processes, the participants oriented themselves to the moderation task, made observations, considered and evaluated candidates' work, made explicit comparisons, formed overall evaluations and a moderation decision, and reflected on this decision. All of the participants made

observations and considerations about multiple candidates within the first ten minutes of moderation, and throughout the investigation stage.

As stated earlier, we compared our data and timelines to the models of Crisp (2017) and Cuff (2017) and we found these models did not fully or accurately describe the observed moderation. Certain aspects of the models, however, were evident in the observed moderation of artwork. This was either directly, or when a different but parallel process was substituted; for example, the processes of scanning and reading in the Crisp (2017) and Cuff (2017) models could be replaced by qualitatively different forms of looking at candidates' artwork (e.g., leafing and consideration).

We found the sequence of activity in the observed moderation sessions most closely resembled the structure of the Crisp (2017) model. In contrast, the specific cognitive processes identified in the Cuff (2017) model more precisely described the subprocesses we observed in the participants' moderation activities. Not every process identified by Cuff (2017) was included in the art moderation model, and we found observed processes in the think aloud data that Cuff (2017) had not included that needed to be added. For example, the participants in this study started comparing submissions from the start of the moderation session.

Similarly to the Cuff (2017) model, we incorporated resources into the art moderation model. We found that all participants used the additional resources, particularly making reference to the standardisation booklet and marking criteria. This is in contrast to Cuff (2017), who found that some moderators relied solely on internalised standards.

We found some differences to both the Crisp (2017) and Cuff (2017) models. Firstly, the participants in this study moved through different candidates' work repeatedly in a cyclical fashion, building up an impression of submission quality. We hypothesise that this was because the work was on full view and did not require detailed or lengthy reading in order to get an impression of quality. Secondly the participants in this study made little, if any, mention of the teacher annotations, whereas this was prominent in the previous research. This could be due to the moderators not needing to read long passages of text and so not needing the hints provided by the annotations.

As the complex structure of both our art process model and the Cuff (2017), model highlights, the moderation process combines many cognitive processes and draws on many resources. That we needed to develop a new model, rather than use one of the existing ones, could be due to the nature of the submissions. The Crisp (2017) and Cuff (2017) models were developed for the moderation of written work, whereas our study used artwork which contained very little text. The overall moderation process should be the same for any subject overseen by the JCQ as it is subject to agreed and documented procedures. Indeed, we found this to be the case for Art and Design, and the overall sequence followed that described by Crisp (2017). We found, however, that the subelements and their interaction with resources did differ for art, suggesting that subject-specific differences may exist at the sublevels.

What all three models emphasise, however, is the iterative and evaluative nature of the moderation process and its focus on quality control. Similarly to Cuff (2017), we found that aspects of the findings support the validity of the moderation process. We too found that the participants followed similar stages, made reference to the marking criteria, focused on appropriate features in the submissions, were mindful of being fair to candidates, and displayed thoroughness in making their judgements.

References

Crisp, V. (2017). The judgement processes involved in the moderation of teacher-assessed projects. *Oxford Review of Education, 43*(1), 19–37.

Cuff, B. (2017). *An exploratory investigation into how moderators of non-examined assessments make their judgements*. (Ofqual/17/6252). Coventry: Ofqual.

Eça, T. (2002). *A conceptual framework for art and design external assessment*. Paper presented at the European Conference on Educational Research, University of Lisbon, Portugal.

Gill, T. (2015). The moderation of coursework and controlled assessment: A summary. *Research Matters: A Cambridge Assessment publication, 19*, 2–6.

Greatorex, J., & Suto, W. M. I. (2008). What do GCSE examiners think of "thinking aloud"? Findings from an exploratory study. *Educational Research, 50*(4), 319–331.

Gruber, D. D., & Hobbs, J. A. (2002). Historical Analysis of Assessment in Art Education. *Art Education, 55*(6), 12–17.

Haanstra, F., Damen, M.-L., Groenendijk, T., & van Boxtel, C. (2015). A review of assessment instruments in arts education. In S. Schonmann (Ed.), *International Yearbook for Research in Arts Education. The wisdom of the many: Key issues in arts education* (pp. 413–418). New York: Waxman.

Herpin, S. A., Washington, A. Q., & Li, J. (2011). *Improving the Assessment of Student Learning in the Arts – State of the Field and Recommendations*. Washington: The National Endowment for the Arts.

Joint Council for Qualifications. (2018). *Instructions for conducting non-examination assessments (new GCE & GCSE specifications)*. Retrieved from https://www.jcq.org.uk/exams-office/non-examination-assessments/instructions-for-conducting-non-examination-assessments

Mason, R., Steers, J., Bedford, D., & McCabe, C. (2005). The effect of formal assessment on secondary school Art and Design education: a systematic description of empirical studies. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

TechSmith. (2011). Morae Recorder [Computer software]. Retrieved from https://www.techsmith.com/morae-features.html.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London, UK: Academic Press.

VERBI Software. (2017). MAXQDA: Software for qualitative data analysis [Computer software]. Retrieved from https://www.maxqda.com/.

What makes researchers anxious? It's *Time to Talk* about talking about research

Gill Elliott, Irenka Suto and **Emma Walland** Research Division

Introduction

Employers participate in mental health initiatives in a number of ways. Key amongst these is the provision of accessible places and events for employees to discuss mental health issues. In the course of discussions in our organisation, Cambridge Assessment, researchers raised the issue of workplace activities that cause them anxiety – top of the list was giving presentations.

Inspired by these conversations, we held a workshop for staff to encourage the discussion to develop more widely, facilitated by the first two authors of this article. The workshop was timed to coincide with the UK national *Time to Talk Day*[1] in 2018 and it was agreed with participants at the outset that a summary of the discussions held during the event would be written up to form this article.

Background

Workplace anxieties have been investigated in the context of more general anxiety disorders, with Linden and Muschalla (2007) finding high levels of work-related anxiety amongst patients admitted to hospital as a consequence of stress. Colligan and Higgins (2006) explored the effects of workplace anxiety upon an organisation and found consequences including low morale, decreased productivity, and increased absenteeism.

The literature around the cognitive and physiological symptoms arising from workplace anxiety is extensive and references poor concentration, rapid heart rate, tremor, sweating, and dry mouth (Kelly & Seveanu, 2005; Wesner, Noyes & Davis 1990). Physiological responses have been linked to the activation of the *fight-or-flight* response of the sympathetic nervous system in response to stress (Harvard Health Publishing, 2018).

The Royal Academy of Dramatic Art (RADA) has an interest in anxiety arising in the workplace. *RADA Business* runs courses in managing nerves and dealing with stage fright, which it extends to a business context:

> *Many people in business experience a workplace equivalent to stage fright. We're calling the nerves, uncertainty and fear that we may feel in workplace scenarios in which we need to perform 'workplace performance anxiety', and many people in their daily working lives experience the feelings associated with it.*
>
> RADA Business (undated)

As a part of its work, RADA Business surveyed 1,000 business people and identified specific situations where workplace performance anxiety was experienced by a proportion of those surveyed. These included job interviews (37.5%), pay negotiations (30.0%), dealing with disagreements (28.1%), giving a presentation (27.5%), making a pitch (24.4%), and networking events (24.2%).

Workplace anxieties which are particularly relevant to the specific work carried out by research staff are those relating to giving presentations and networking, including the 'cold calling' via telephone which is sometimes needed to recruit project participants. RADA's description of workplace performance anxiety sounds very similar to the anxiety that researchers have talked about in conversations about giving presentations.

In this article, we use the terms *presentation*, *performance*, and *workplace anxiety*. They are not necessarily interchangeable as they refer to an increasingly broad set of situations, but are each used in the context to which they seem most applicable.

The workshop

The hour-long workshop was held on 1 February, 2018 – the designated national *Time to Talk Day*. The full title (shown as Figure 1) reflected one of the key activities in a researcher's role which tends to cause anxiety – giving presentations.

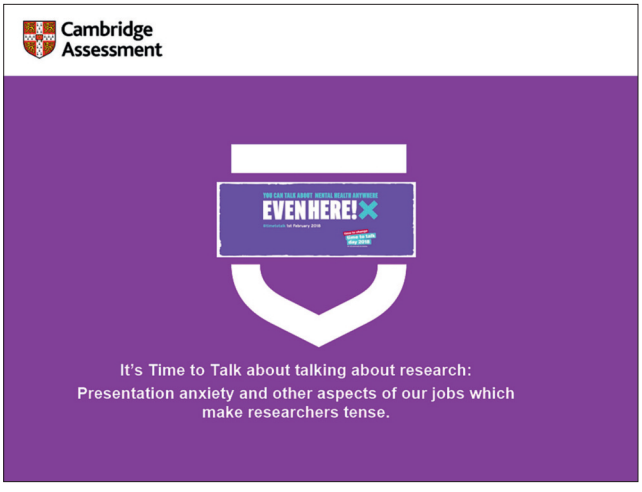


Figure 1: Title of the workshop

The workshop also allowed for the additional exploration of other sources of anxiety connected with talking about research, such as making phone calls to potential research participants, and discussing research performance in annual performance reviews. The key aim of the workshop was to start an open conversation amongst researchers at Cambridge Assessment. This conversation could have a wider impact by prompting other colleagues across the Cambridge Assessment Group, or in other organisations, to reflect on their experiences of workplace anxiety.

The format of the workshop included a brief introduction to the topic, followed by a 20-minute discussion in table-based groups about when,

1. *Time to Talk Day* is coordinated by *Time to Change*, a social movement working to change the way we all think and act about mental health. It says: "On *Time to Talk Day*, we want everyone to have a conversation about mental health". *Time to Change* is led by *Mind and Rethink Mental Illness*.

how, and what anxiety was experienced by participants in the course of their normal work. This was followed by a short presentation of key elements of the literature. Finally, a second round of open discussion was held in the same table groups as before, brainstorming practical strategies for reducing the impact of performance anxiety.

Table discussions format

Fifteen researchers from Cambridge Assessment attended the event. The participants represented a range of job descriptions and levels. Participants were free to sit where they chose, naturally ordering themselves into four small groups of approximately four participants each. This ensured a relatively random allocation to the groups, although in some cases friends or close colleagues will have chosen to sit together.

Table discussions guidelines

After the general introduction, each group was given 5 stimulus questions to discuss for 20 minutes. One member of each group was asked to take notes under the headings of the questions, and these notes formed the data corpus for analysis. The order in which the questions were presented to each group was intentionally varied to reduce the impact of adjacent tables' conversations, and to ensure that, if time ran out, all questions would have been addressed. The stimulus questions were:

1. Which workplace occasions prompt performance anxiety?
2. Are we open with colleagues if we are nervous? If not, why is that?
3. What are the physical symptoms of presentation anxiety and when do they occur?
4. What are the emotional symptoms of presentation anxiety and when do they occur?
5. Are there any other aspects of the issue you discussed?

The nature of anxiety experienced by workshop participants

The notes from the table discussions were analysed in three main stages. These broadly adhered to the principles of thematic analysis according to Braun and Clarke (2006), although a full-scale thematic analysis was not attempted due to the limited nature of the data. The first stage of analysis consisted of familiarisation with the responses through repeated reading. Secondly, common themes across the responses were identified. Thirdly, the themes were refined and, lastly, they were written up.

The themes yielded by analysis fell under three broad categories, which were aligned with the discussion questions: (a) openness about anxiety, (b) sources of anxiety, and (c) implications of anxiety. Table 1 shows the identified themes under their respective categories.

Table 1: Categories and themes about workplace anxiety identified by workshop participants

Category	*Openness about anxiety*	*Sources of anxiety*	*Implications of anxiety*
Themes	Role of culture	In the spotlight	Socio-emotional
	Workplace relationships	Conflict	Physiological
	Anticipated negative consequences	Hierarchy	Cognitive
		Appearing unintelligent	
		Uncertainty	

The following paragraphs illustrate each theme using extracts that supported the data. The extracts could be either individual opinions of group members, or opinions formed by group consensus. Since we were unable to determine this with certainty, we have not attempted to make this distinction in the findings. The groups are referred to as Group 1 to 4.

Openness about anxiety

Participants across all groups reported some circumstances in which they were open about their anxieties, and others in which they were not. Their openness about anxiety was influenced by cultural factors, workplace relationships, and anticipated consequences.

The role of culture

All four groups made reference to the role of culture in their openness about anxiety. Group 4 commented that talking about anxiety was not part of the organisational culture – "not much culture in the office about talking (generally) and this kind of thing". Group 1 noted that they would "share when it's something it's socially acceptable to be nervous about". This points to the role cultural norms have in determining whether anxiety is acceptable or not. Role expectations also featured. For example, Group 2 noted a reason they were not open about anxiety was that they "want to meet expectations", and Group 3 would not be open due to a "fear of letting the side down". However, they also noted they would share if they were invited to, indicating that employees may be open to talking about anxiety if this is encouraged by their colleagues. Timing may be important here, as Group 1 noted they might be more likely to talk about it after the event.

The role of workplace relationships

Workplace relationships featured prominently in discussions on openness about anxiety. Participants reflected on how their openness varied depending on whom in the organisation they were talking to, and what their relationship with that person was. For example, Group 2 noted "It depends on the colleague"; Group 3 wrote "[It] depends on both the personal and professional relationship you have with them"; Group 1 noted "[We would] share with people we are comfortable with"; and Group 4 said "[We would be open about our anxiety] if [we] feel close/comfortable with someone". The extracts suggest that the strength of the relationship was an important factor in openness. If there was a weaker relationship, people were less willing to share: For example, Group 3 said that the "level of existing workplace relationship [is] not yet ready".

For Group 2, relationships and openness were related to the organisational structure. They noted being less willing to share anxiety with colleagues from another department (known as 'division' in our organisation), but being more willing to share with their close team. They wrote "More likely to share within the team – save divisional face". This may be a reflection of the relative strength of relationships of different teams. It is interesting that this group noted that sharing anxieties with team members allowed them to "save divisional face". The team possibly provides a safe place for sharing anxieties, whilst avoiding embarrassment at a divisional or organisational level.

Anticipated negative consequences of being open about workplace anxiety

All groups reflected on the possible negative consequences of openness about anxiety. Negative consequences for the most part related to the potential for reputational damage. Group 2's note of "save divisional face" suggests that openness about anxiety could lead to humiliation. They also noted "people think worse of you and want to save face". Group 3 echoed this sentiment with "Fear of being judged", and Group 1 said "We're trying to avoid looking stupid". Conversely, some positive consequences were reported; for example, Group 4 noted that telling others about anxiety "gives an excuse". Caution should be exercised when interpreting these findings as it is unclear whether these anticipated negative consequences are based on actual past experiences, or if they are hypothetical or imagined fears.

Sources of anxiety

Analysis of the data yielded five themes relating to sources of anxiety, which were: (a) work being placed in the spotlight, (b) appearing or feeling unintelligent, (c) conflict, (d) hierarchy, and (e) uncertainty.

Although the themes are presented separately for theoretical purposes, it must be noted that each data extract may reflect various themes. For example, "receiving feedback on reports" could be a source of anxiety due to four of the above five themes:

- Appearing/feeling unintelligent – if the feedback is negative;
- Conflict – possible disagreement about the feedback;
- Hierarchy – if feedback is given within the context of an unequal power relationship; and
- Uncertainty – the element of the unknown related to what feedback will be received.

In the spotlight

The majority of reported sources of anxiety across all of the four groups related to workplace events in which work was placed in the spotlight for possible scrutiny by others, whether internally or externally. Commonly reported sources included "giving presentations"; "giving a talk"; "making and receiving phone calls, especially if overheard (e.g., in an open plan office)"; "sending drafts for people to see"; "being videoed"; "feedback meetings with advisors"; "receiving feedback on reports"; "annual performance appraisal meetings"; "report review meetings"; and "being interviewed". A specific instance of performing in front of others with high potential for embarrassment was "using technology in front of people with no practice (e.g., a conference call)".

Some sources of anxiety for Groups 3 and 4 were situations where work was under the spotlight to external parties. These included "feedback meetings with customers", "chairing conference sessions", "sending external emails", "publishing", and "talking to journalists". Participants reported pressure related to representing the organisation in such situations (e.g., "sending external emails").

Appearing/feeling unintelligent

A few sources of anxiety for Groups 1 and 3 seemed to reflect a fear of appearing or feeling less intelligent in front of others. For example, "asking about something you probably should know (feeling stupid)" and "feedback meetings with certain people perceived as less tolerant of low knowledge". Group 3 also noted *imposter syndrome* as a source of anxiety, which refers to a belief that one is unintelligent despite numerous achievements and other evidence to the contrary (Clance & Imes, 1978). "Presenting someone else's work (that you might not even agree with)" could also be an instance of this theme, although the anxiety could result from other factors, such as a lack of confidence in the material to be presented.

Conflict

For Groups 1 and 2, a few sources of anxiety were the result of situations in which conflict was anticipated. These included "meetings with difficult colleagues", "asking someone to do something", "difficult phone calls", "meetings where you anticipate conflict", "meetings where you have to give negative feedback", and "meetings where you have to pass on bad or unwelcome news".

Hierarchy

This theme was prevalent across all four of the groups, and refers to sources of anxiety that were related to positions of power within the organisation. For example, "meetings with senior people", "meetings with high-status colleagues", "meetings with advisors", and "report review meetings" were all noted as sources of anxiety. "Asking someone to do something", which was mentioned in the previous theme, could also provoke anxiety due to power dynamics.

Uncertainty

For Groups 1 and 4, situations in which there was a high degree of uncertainty were noted to provoke anxiety. These were: "calling someone we don't know" and "[taking part in] calls with unknown people, especially conference calls" respectively. Group 2 also noted in their discussions about the symptoms of anxiety and "fear of the unknown".

Implications of anxiety on a researcher's functioning

All groups recognised the negative impact that anxiety had on their socio-emotional, physiological, and cognitive functioning.

Socio-emotional

Various negative emotions were associated with anxiety, including "fear", "panic", "terror", "annoyance", "irritability", "less happy" (low mood state), and "feeling small". Beyond these transitory emotions, anxiety was noted among some groups to have a wider impact on their lives through affecting their interpersonal relationships. Group 3 noted "being intolerant – reduced capacity to engage" and Group 1 noted that it could "affect inter-personal relationships".

Physiological

Anxiety was perceived to cause a myriad of negative physiological changes. These included symptoms associated with activation of the sympathetic nervous system which include gastrointestinal complaints ("stomach ache", "needing the loo", and "digestive problems"); "sweating"; speech production issues ("false start to sentences", "stumbling over words", "shaky voice", and "speaking less fluently,

faster or more quietly"); "shaking"; "increased heart rate"; "sleep problems"; and "dry mouth".

Cognitive

Although the groups were not specifically asked to discuss the cognitive symptoms of anxiety, a few groups made mention of them. Group 1 noted *catastrophising*, which refers to focusing on the worst possible outcome, however unlikely, or experiencing a situation as unbearable or impossible when it is just uncomfortable. Groups 1, 2 and 3 made reference to the cognitive process of *rumination* (*negative replay* and *ruminating*). Rumination refers to going over a negative thought or problem without activating the process of problem solving. Furthermore, Group 3 noted how anxiety can cause 'distraction-can't focus', which refers to an impairment in concentration.

In summary, as expected, the sources of anxiety were predominantly performance related, in situations where participants felt they were in the spotlight. The participants sometimes feared possible negative evaluation or judgement. However, not all their anxieties were strictly related to traditional performance situations. Imposter syndrome was noted as a source of anxiety and there were a few interesting cases of anxiety in situations with high potential for conflict, and situations in which power imbalances were present. The latter two could relate to the broader construct of social anxiety, but they may also relate to fears of the unknown, and feelings of loss of control.

In terms of openness about anxiety, participants reflected on the different circumstances in which they would or would not be open about anxiety. The analysis pointed to the role that culture plays in this, specifically in terms of cultural norms and role expectations. Lastly, there was a wide range of socio-emotional, physiological, and cognitive symptoms occurring with anxiety.

Relating our researchers' reported experiences to the literature

Participants' experiences of social and performance anxiety in the workshop were similar to symptoms of social anxiety in general, and performance anxiety in particular. The anxiety seemed to revolve around a fear of negative evaluation by others, such as the fear of looking unintelligent or potentially embarrassing oneself. These findings were in line with what was expected, since the workshop was particularly focused on performance anxiety. While most of the anxiety-provoking situations in this research were performance-only, some could be related to the wider construct of social anxiety, such as talking to an unknown person on the phone.

Anxiety is both an emotional and physical condition, and therefore results in many symptoms across the socio-emotional, cognitive, and physiological domains. The symptoms reported by participants in this research were consistent with known effects of anxiety (American Psychiatric Association, 2013). The participants' discussions around cultural elements of anxiety revealed an awareness of the way in which cultural norms and expectations influenced their willingness to talk about anxiety. In particular, social norms were important as they noted how some forms of anxiety are considered acceptable, but not others.

Role expectations also influenced openness about anxiety. Role expectations reflect what is expected of individuals who occupy certain socially defined categories (Parsons, 1991). Participants in this research referred to expectations and standards they felt under pressure to conform to, and which, at times, limited their openness. Hierarchy can also play a part. For example, situations characterised by unequal power relationships can provoke anxiety due to a lack of control for the individual occupying the lower status position (Gray, 1991; Raghunathan & Pham, 1999; Smith & Ellsworth, 1985).

Participants were concerned about looking and feeling unintelligent, a core feature of imposter syndrome. This syndrome also has cultural dimensions: experienced psychologists have noted that imposter syndrome is more likely to occur in cultures that place high value on achievement (Weir, 2018). The fear of appearing unintelligent was a source of anxiety, as well as a factor limiting openness about anxiety.

Participants' identification of uncertainty and fear of the unknown were not unexpected. Uncertainty and lack of control can be seen as two major fears underlying many forms of anxiety (Carleton, 2016; Gray, 1991; Raghunathan & Pham, 1999; Smith & Ellsworth, 1985), and they may explain why conflictual situations, and situations involving power imbalances, were perceived by participants to provoke anxiety. Conflicts are generally unpredictable and are often not in our control, and situations of power imbalances also result in a lack of control for the party with less power.

Respondents' reflections on the cognitive action of catastrophising when faced with certain situations are not surprising. Catastrophising is linked with performance anxiety in particular; for example, it has been found among students in relation to taking their GCSE examinations (Putwain, Connors, & Symes, 2010). Rumination is more concerning: it is a specific type of repetitive negative thinking (McLaughlin & Nolen-Hoeksema, 2011) and is a common factor in both anxiety and depression (Harrington & Blankenship, 2002).

Research on reducing performance anxiety

Research on reducing performance anxiety has come a long way in recent years. In our view, the technique underpinned by the strongest evidence base is that developed by Brooks (2014). Her key insight is to get people excited rather than nervous. Brooks argues that in psychological terms, anxiety can be conceptualised as an emotion characterised by high physiological arousal and negative valence, as well as uncertainty and a low sense of control (Gray, 1991; Raghunathan & Pham, 1999). This point is illustrated in Table 2, which also contains some contrasting emotions.

Table 2: Anxiety as an emotion

	High physiological arousal	*Low physiological arousal*
Positive valence (mood)	Excitement	Calmness
Negative valence (mood)	Anxiety	Boredom

A substantial literature indicates that reappraising negative emotions is more effective than suppressing them (e.g., Gross, 1998, 2001; Gross & Levenson, 1993; Hofmann, Herring, Sawyer, & Asnaani, 2009). Therefore, it might make sense superficially to attempt to reduce anxiety by reappraising it as calmness; that is, by telling ourselves we feel calm. Brooks argues, however, that calming ourselves down can be very difficult because it entails crossing both of the axes in Table 2. During high physiological arousal, we experience surges of adrenaline, causing

our hearts to race and our stomachs to turn over, and these processes are not easy to control consciously. Instead, it is easier to reduce anxiety by reappraising it as excitement. This entails crossing only the valence axis in Table 2. That is, it involves accepting a state of high physiological arousal but making ourselves feel positive rather than negative. We can move from anxiety to excitement by visualising something positive, and/or finding a reason to be excited about the impending task.

To test her ideas, Brooks (2014) conducted a series of experiments. In the first of these, 113 participants were informed that they would soon have to sing a popular song in public. In response to this news, their heart rates rose considerably. Prior to their singing performances, participants read out one of three statements:

1. *I am excited;*
2. *I am anxious;* or
3. *I am calm.*

Voice recognition software was used to assess each participant's singing performance in terms of pitch, volume, and rhythm. The singing performances of 'excited' performers scored an average of 81%. In contrast, 'anxious' performers scored an average of 69%, and 'calm' performers scored an average of 53%.

Brooks (2014) repeated her experiment using a public speaking task. Each of 140 participants had to give a 3-minute presentation explaining why they would be good work colleagues. Participants who said they were excited scored higher across measures of persuasiveness, competence, confidence, and persistence compared to participants who said they were anxious or who explicitly attempted to calm themselves down. Similar results were achieved in a study of 188 participants, each of whom attempted a difficult mathematics problem (Brooks, 2014). Overall, Brooks' findings provide a convincing demonstration that, prior to undertaking anxiety-inducing tasks, the way we talk about our feelings influences whether we feel anxious or excited, and this has a significant influence on our subsequent performance.

Summary of notes recorded by groups on practical techniques for reducing the impact of anxiety

The final session of the workshop was a brainstorming discussion about techniques that participants knew about, or had used to help ameliorate anxiety. The specific question which was presented as a stimulus was: *What other strategies are there for handling anxiety?* Figure 2 shows a completed recording sheet from one of the groups.

Table 3 presents the complete list of techniques that were generated during the session, in the participants' own words. In each case, an indication is presented of whether the technique is used before or during the event.

Nearly all of the techniques noted by participants in the final brainstorming session relate specifically to presentations, even though the stimulus prompt was more general. This may have occurred as a result of this being the overriding concern for many participants, or it may have been that the presentation from the literature tended to lead them to this point. The techniques suggested by workshop participants relate back to issues identified earlier in the workshop.

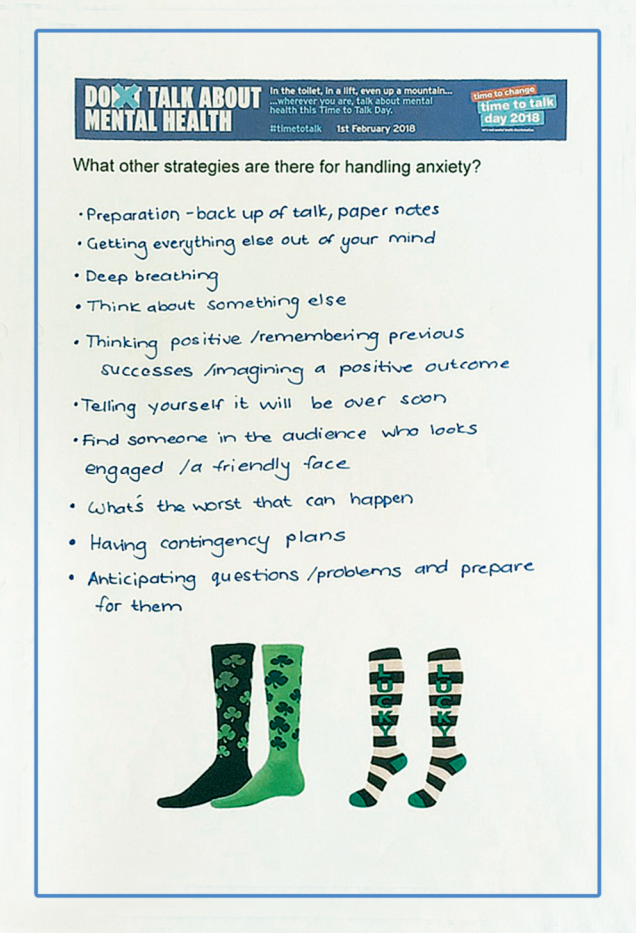


Figure 2: Completed recording sheet from the final discussion session

Social and performance anxiety

Suggestions of techniques to alleviate fears of negative evaluation or being embarrassed included preparation, thinking positively/imagining a positive outcome, finding a friendly face in the audience, focusing upon the opportunity with positive benefits to be gained, and having a presentation persona.

The technique of preparing for questions and anticipating problems ahead of time suggested in the final discussion session is one which directly addresses the issue of being perceived as unintelligent. More importantly, the knowledge of that preparation helps the individual to feel that they are not going to be in danger of being caught out.

Uncertainty and loss of control

Participants engaged with the need to have tactics for dealing with feelings of a loss of control and with potential conflict. As well as preparation, already discussed in the context of culture, participants suggested arriving early to test the presentation. This avoids being caught out by unexpected technical problems which might place the speaker in unfamiliar territory. Having anchor points within the presentation acts in a similar way – placing extra control back into the presenter's hands.

Socio-emotional, cognitive, and physiological responses

Many of the suggestions made by participants were in response to alleviating physiological symptoms of anxiety: breathing exercises, singing, and physical activity to dissipate energy. Some interesting suggestions were made which relate to managing cognitive elements of

Table 3: Techniques for alleviating anxiety in presentations and similar events, as described by workshop participants

	In advance	*Just beforehand*	*During*
Preparation (paper notes, practice – particularly the beginning, check the room)	◆		
Have contingency plans (back-up of talk, paper notes)	◆		
Getting everything else out of your mind		◆	◆
Think about something else	◆	◆	
Breathing exercises	◆	◆	◆
"Thinking positive" by remembering previous successes or imagining a positive outcome	◆	◆	
Telling yourself it will be over soon		◆	◆
Find someone in the audience who looks engaged/a friendly face			◆
Think: What's the worst that can happen? Remember the grand scheme of things. Nobody will die. It's highly unlikely that you'll be the worst presenter they've ever seen.	◆	◆	
Anticipating questions/problems and preparing for them	◆		
Arrive early and test the presentation		◆	
Sleeping tablets		◆	
Wear comfortable/"confident" clothing		◆	◆
Have a drink		◆	◆
Sing a happy song		◆	
Practise a good physical stance	◆		
Run up the stairs to use up adrenaline		◆	
Tell a joke			◆
Lucky underwear – a fun secret that reminds you that you have a life outside		◆	◆
See the event as an opportunity with everything to gain	◆	◆	◆
Have a presentation persona – a professional sort of separate self		◆	◆
Try not to be alone so you don't immediately have dwelling time	◆	◆	
Have anchor points within the presentation	◆		◆

anxiety. The presenters had noted suggestions such as "focusing upon a nice meal or activity as a self-reward after the event". Additional suggestions from workshop participants included two strategies for refocusing perspective: (a) "telling yourself it would be over soon", and (b) "wearing particular clothes as a confidence booster and reminder of life beyond work".

Openness and cultural elements of anxiety

Participants reported being open about their anxiety to colleagues in their teams with whom they had strong relationships. This is a protective factor since social support reduces the negative impact that anxiety has on a person's functioning. Cultivating strong workplace relationships may be a productive step for organisations to take to increase openness.

Limitations of the research

The research has several limitations. Firstly, the group who participated were self-selected and the workshop might reasonably be assumed to have not attracted people who are totally unaffected by anxiety, or those who are so badly affected that they prefer not to engage in discussions about it. Due to the cultural influences on openness around anxiety, participants may have been reluctant to share their true feelings in front of their colleagues and they therefore may have only discussed socially acceptable anxieties. This means that there could be a greater number of anxiety-provoking workplace situations that were not revealed in this research. There is also the possibility with roundtable discussions that one or a few participants dominated the discussion and the views of some participants were not represented. Errors could have been introduced as the person taking the notes may not have accurately recorded what others meant, and there is a possibility of misinterpretation of the notes during analysis. There were some instances where the meaning in what was written was not entirely clear and more detail would have aided analysis. Lastly, the wording of the questions led participants to focus on performance anxiety; therefore, few other sources of anxiety in the workplace were uncovered.

Conclusion

Core to this research was how workplace situations led to performance and social anxiety as experienced by Cambridge Assessment researchers who participated in the workshop:

- Participants feared negative evaluation and humiliation by others.
- Traditional performance situations provoked anxiety, but conflict situations and situations involving hierarchy were also noted as sources of anxiety.
- Participants were sometimes not open with others about their anxiety due to the fear of judgement, and this illuminated the role of cultural norms and expectations in openness about mental health issues, even in an organisation which takes great care to encourage this.
- Elements of culture and workplace relationships can either encourage or discourage behaviours, including talking about mental health.
- Participants were open about their anxiety if they considered it to be socially acceptable, and if they had strong relationships with their team members.
- Participants' anxieties were accompanied by a wide range of physiological, socio-emotional, and cognitive responses, all of which are known to accompany anxiety.

Organisations need to be aware of the situations which provoke anxiety amongst employees in order to enable the provision of support

and intervention. We anticipate that this research will be helpful in enabling research staff in our, and other, organisations to become more open in discussing the aspects of their role which cause anxiety. In turn, this will facilitate finding ameliorative solutions to the issues raised, to the mutual benefit of individuals and organisations. We hope that we have achieved the aim of opening a discussion among our own researchers and, with this article, we hope to extend that discussion to a wider audience.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Washington, DC: American Psychiatric Association.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

Brooks, A. W. (2014). Get Excited: Reappraising Pre-Performance Anxiety as Excitement. *Journal of Experimental Psychology: General, 143*(3), 1144–1158.

Carleton, R. N. (2016). Fear of the unknown: One fear to rule them all? *Journal of Anxiety Disorders, 41*, 5–21.

Clance, P. R., & Imes, S. A. (1978). The imposter phenomenon in high achieving women: Dynamics and therapeutic intervention. *Psychotherapy: Theory, Research & Practice, 15*(3), 241.

Colligan, T., & Higgins, E. (2006). Workplace Stress. *Journal of Workplace Behavioral Health, 21*(2), 89–97.

Gray, J. (1991). Fear, panic, and anxiety: What's in a name? *Psychological Inquiry, 2*(1), 77–78.

Gross, J. J. (1998). Antecedent- and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology. *Journal of Personality and Social Psychology, 74*(1), 224–237.

Gross, J. J. (2001). Emotion regulation in adulthood: Timing is everything. *Current Directions in Psychological Science, 10*(6), 214–219.

Gross, J. J., & Levenson, R. W. (1993). Emotional suppression: Physiology, self-report, and expressive behavior. *Journal of Personality and Social Psychology, 64*(6), 970–986.

Harrington, J., & Blankenship, V. (2002). Ruminative thoughts and their relation to depression and anxiety. *Journal of Applied Social Psychology, 32*(3), 465–485.

Harvard Health Publishing (2018). *Understanding the stress response*. Retrieved 20 June, 2018, from https://www.health.harvard.edu/staying-healthy/understanding-the-stress-response

Hofmann, S. G., Heering, S., Sawyer, A., & Asnaani, A. (2009). How to handle anxiety: The effects of reappraisal, acceptance, and suppression strategies on anxious arousal. *Behaviour Research and Therapy, 47*(5), 389–394.

Kelly, V. C., & Saveanu, R. V. (2005). Performance anxiety: How to ease stage fright. *Current Psychiatry, 4*(6), 25–34.

Linden, M., & Muschalla, B. (2007). Anxiety disorders and workplace-related anxieties. *Journal of Anxiety Disorders, 21*(3), 467–474.

McLaughlin, K. A., & Nolen-Hoeksema, S. (2011). Rumination as a transdiagnostic factor in depression and anxiety. *Behaviour Research and Therapy, 49*(3), 186–193.

Parsons, T. (1991). *The social system*. London: Routledge.

Putwain, D. W., Connors, L., & Symes, W. (2010). Do cognitive distortions mediate the test anxiety–examination performance relationship? *Educational Psychology, 30*(1), 11–26.

RADA Business (undated). *Beating workplace performance anxiety*. Retrieved 2 July, 2018, from https://www.radabusiness.com/about-us/research/

Raghunathan, R., & Pham, M. (1999). All negative moods are not equal: Motivational influences of anxiety and sadness on decision making. *Organizational Behavior and Human Decision Processes, 79*(1), 56–77.

Smith, C., & Ellsworth, P. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology, 48*(4), 813–838.

Weir, K. (2018). *Feel like a fraud?* Retrieved 20 June, 2018, from http://www.apa.org/gradpsych/2013/11/fraud.aspx

Wesner, R., Noyes Jr, R., & Davis, L. (1990). The occurrence of performance anxiety amongst musicians. *Journal of Affective Disorders, 18*(3), 177–185.

Cambridge Assessment Network

Leadership in National Assessment

"This course is a great overview of the assessment industry around the world – its problems, challenges, trends and future directions. This type of knowledge is fundamental for someone leading a modern and open assessment agency."

2017 LNA course delegate

Are you a senior leader in assessment?

Join us for *Leadership in National Assessment* this September: a unique programme providing a transformative opportunity to learn from a range of acclaimed experts and to share your own knowledge and expertise with a multinational group of peers.

The week-long course is designed for delegates from around the world and covers:

- The role of large-scale assessment organisations and their place in the wider education context
- How is, and how should, education policy be decided? The role of evidence-based practice
- A Systems Approach to managing this complexity.

16 – 20 September 2019 | Cambridge, UK

Find out more: www.canetwork.org.uk/leadership

Research News

Karen Barden Research Division

Conference presentations

British Educational Research Association (BERA)

In September 2018, the British Educational Research Association (BERA) went to Northumbria University in Newcastle upon Tyne, UK, to host its 44th conference. Each year, the conference provides a valuable opportunity for researchers to share their work and to discuss and debate across many educational themes. Several researchers from Cambridge Assessment attended the conference and the following papers were presented:

Filio Constantinou and Lucy Chambers, Research Division: *'That path won't lead nowhere': non-standard English in UK students' writing over time.*

Simon Child, Cambridge Assessment Network, and Sylvia Vitello, Research Division: *Faster, higher, stronger? A comparison of content coverage and teacher pedagogy in Technical Awards and GCSEs in sport-related subjects.*

Tim Gill, Research Division: *How have students and schools performed on the Progress 8 performance measure?* The presentation and commentary can be viewed as a *Research Byte* at www.youtube.com/watch?v=0qEtOfJAiMs

International Association for Educational Assessment (IAEA)

The 44th Annual Conference of the IAEA took place in Oxford, UK, in September 2018 with the theme of *Assessment and Big Data*. The conference focused on new ways of discovering and quantifying important Big Data measures with sufficient reliability and validity, and new quantitative techniques that will emerge from the world of Big Data. Several researchers from Cambridge Assessment attended the conference and the following papers were presented:

Matthew Carroll, Research Division: *News articles as data: analysing the portrayal of exams in the UK print media*. The presentation and commentary can be viewed as a *Research Byte* at www.youtube.com/watch?v=KZGjWVso9Kw

Tom Sutch, Nicole Klir, and Anne-Claire Gueranger, Research Division: *Gaining insight from social media data in educational assessment.* The presentation and commentary can be viewed as a *Research Byte* at www.youtube.com/watch?v=yOOkPSvLf0M

Martin Johnson, Research Division, and Stuart Shaw, Cambridge Assessment International Education: *Developing a framework for assessing computer-based test washback*.

Edmund Jones and Rick Chalton: Cambridge Assessment English: *Using Big Data to counteract cheating on tests*.

Sarah Mattey and Stuart Shaw, Cambridge Assessment International Education: *Predictive validity: Using success on acceleration programmes for postsecondary admissions and predicting university success*.

Nick Raikes, Research Division: *Data, data everywhere? Opportunities and challenges in a data-rich world*. A transcript of the presentation features as the third article in this issue on pp.16–19. The presentation and commentary can be viewed as a *Research Byte* at www.youtube.com/watch?v=8_FP6YDDJ1I&t=2s

Nicky Rushton, Research Division: *What lessons from current working practice can be applied to big data? Identifying GCSE equivalents across many jurisdictions*.

Carmen Vidal Rodeiro, Research Division, and Stuart Shaw: *The value of predictive validity studies and the need for 'fit-for-purpose' data to inform postsecondary admissions policies and decision-making in the United States*.

Stuart Shaw: *The construction of a validity portfolio for general educational qualifications: A suggested approach to large-scale validation.*

Stuart Shaw and Marta Czajkowska, Cambridge Assessment International Education: *Research issues and challenges in the design, construction and implementation of an L2 school student survey.*

Sylvia Vitello and Joanna Williamson, Research Division: *Effects of AS/A level reforms in England: Extra insights from the National Pupil Database.*

Wafa Al-Yaqoobi, Basma Al-Sedeq, Directorate of National Examinations, National Authority for Qualifications and Quality Assurance of Education and Training, Bahrain, and Stuart Shaw: *Evaluating the impact of the Bahrain National Examinations*

Stuart Shaw also led a pre-conference workshop on *Achieving in Content Through Language: Assessing Bilingual and Multilingual Learners*.

EARLI SIG 14 - Learning and Professional Development

The EARLI Special Interest Group (SIG) 14 - Learning and Professional Development focuses on professional and vocational education and training and has a particular set of interests associated with learning at, and for, work. The 9th edition of the conference took place in Geneva, Switzerland, in September 2018 and was dedicated to the topic of *Interaction, learning and professional development*. Martin Johnson, Research Division, presented a paper on *Learning to think alike: A study of professional examiners' feedback interactions in a UK Qualification Awarding Organisation.*

Association for Educational Assessment-Europe (AEA-Europe)

Held in November 2018, the 19th AEA-Europe Annual Conference took place in Arnhem-Nijmegen, the Netherlands, with the theme of *Building bridges to future educational assessment*. Several researchers from Cambridge Assessment attended the conference and the following papers were presented:

Tom Benton, Research Division: *Exploring the relationship between optimal methods of item scoring and selection and predictive validity.*

Tom Bramley, Research Division: *Evaluating the 'similar items method' for standard maintaining.*

Filio Constantinou, Research Division: *Constructs over time: the construct of language competence as captured in tests in England over the last 150 years.*

Victoria Crisp, Tom Bramley, Research Division, and Stuart Shaw, Cambridge Assessment International Education: *Should we be banking on it? Exploring potential issues in the use of 'item' banking with structured examination questions.*

Martin Johnson, Research Division, and Stuart Shaw: *A Study in Computer-Based Testing Washback: A Framework for Supporting Practitioner Research.*

Isabel Nisbet, Faculty of Education, University of Cambridge, and Stuart Shaw: *"It all depends on how we look at things" – Bringing a lens to fairness in assessment.*

Tim Oates, Assessment Research and Development: *Individualised learning, personalised learning – implications for assessment.*

Sarah Mattey and Stuart Shaw, Cambridge Assessment International Education: *Predictive validity: Using success on acceleration programmes for postsecondary admissions and predicting university success.*

Joanna Williamson and Matthew Carroll, Research Division: *Vocational, technical and applied qualifications within upper secondary education in England: Who takes them, how do they fit within students' programmes of study, and where do students progress to next?* The presentation and commentary can be viewed as a *Research Byte* at www.youtube.com/watch?v=ejgWpNj0RNk

Stuart Shaw also led a pre-conference workshop on *Achieving in Content Through Language: Assessing Bilingual and Multilingual Learners.*

MAXQDA User Conference

The MAXQDA User Conference took place in February 2019 in Berlin, Germany with the theme *Learn. Discover. Connect*. The event centred on questions on how to optimize the use of MAXQDA in the various methodological and thematical settings of qualitative and mixed-methods research. This year is the 30th anniversary of MAXQDA and the welcoming address covered the history, present, and future of MAXQDA. There were a number of spotlight sessions and 'round tables' giving a chance to explore features of the software, and discuss questions and techniques with other delegates.

Lucy Chambers, Research Division, gave a poster presentation on *Moderating artwork: Investigating judgements and cognitive processes.* This was based on her research with colleagues, Joanna Williamson and Simon Child, and features as the fourth article in this issue on pp.19–25.

Further information on all conference papers can be found on our website: www.cambridgeassessment.org.uk/our-research/all-published-resources/conference-papers/

The Cambridge Approach to Curriculum

This event was held in Cambridge, UK, in March 2019. It included presentations from Tim Oates, CBE, Assessment Research and Development, Martin Johnson and Tori Coleman, Research Division, and Abigail Barnett, Cambridge Assessment International Education. Delegates discussed key issues around how to develop a working model of a curriculum, as well as some of the core issues in curriculum development.

UK Rasch Users' Group

Cambridge Assessment hosted the 13th annual meeting of the UK Rasch Users' Group in March 2019 at The Triangle Building in Cambridge, UK. It was attended by over 70 Rasch users from universities and external organisations including Ofqual, AQA, and Pearson UK. The event provided delegates with an informal forum for sharing work and networking, as well as the opportunity to hear presentations including those from Joanna Williamson, Matthew Carroll, and Tom Benton, Research Division. Further insight from the event can be found on our website: www.cambridgeassessment.org.uk/blogs/rasch-researchers-collaborate-at-cambridge/

Publications

The following articles have been published since *Research Matters*, Issue 26:

Bramley, T., & Crisp, V. (2019). Spoilt for choice? Issues around the use and comparability of optional exam questions. *Assessment in Education: Principles, Policy & Practice, 26*(1), 75–90. Available online at DOI: 10.1080/0969594X.2017.1287662

Carroll, M., & Benton, T. (2018). *The link between subject choices and achievement at GCSE and performance in PISA 2015*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at www.cambridgeassessment.org.uk/Images/517876-the-link-between-subject-choices-and-achievement-at-gcse-and-performance-in-pisa-2015.pdf The Executive summary of the report is available at www.cambridgeassessment.org.uk/Images/517878-the-link-between-subject-choices-and-achievement-at-gcse-and-performance-in-pisa-2015-executive-summary.pdf

Constantinou, F., Chambers, L., Zanini, N., & Klir, N. (2019). A diachronic perspective on formality in students' writing: empirical findings from the UK. *Language, Culture and Curriculum*. Advance online publication available at DOI: 10.1080/07908318.2018.1563609

Gill, T. (2018). *The impact of the introduction of Progress 8 on the uptake of qualifications in English schools – an update for 2016/17*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at www.cambridgeassessment.org.uk/Images/

519667-the-impact-of-the-introduction-of-progress-8-on-the-uptake-of-qualifications-in-english-schools-an-update-for-2016-17.pdf

Nisbet, I., & Shaw, S. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy and Practice*. Advance online publication available at DOI: 10.1080/0969594X.2019.1586643

Johnson, M., & Shaw, S. (2018). What is computer-based testing washback, how can it be evaluated, and how can this support practitioner research? *Journal of Further and Higher Education*. Advance online publication available at DOI: 10.1080/0309877X.2018.1471127

Shaw, S., & Vidal Rodeiro, C.L. (2019). The Value of Predictive Validity Studies and the Need for 'Fit-for-Purpose' Data to Inform Postsecondary Admissions Policies and Decision Making. *Strategic Enrollment Management Quarterly,* 6(4), 23–39. Available online at www.aacrao.org/research-publications/quarterly-journals/sem-quarterly/article/volume-6/issue-4/the-value-of-predictive-validity-studies-and-the-need-for-fit-for-purpose-data-to-inform-postsecondary-admissions-policies-and-decision-making-in-the-united-states

Vidal Rodeiro, C.L., & Williamson, J. (2018). Meaningful destinations: using national data to investigate how different education pathways support young people's progression in England. *Research Papers in Education*. Advance online publication available at DOI: 10.1080/02671522.2018.1536889

Walland, E., & Darlington, E. (2019). *Insights on trends in AS Levels, the EPQ and Core Maths: summary report*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at www.cambridgeassessment.org.uk/Images/527125-insights-on-trends-in-as-levels-the-epq-and-core-maths-summary-report.pdf

Williamson, J., & Vitello, S. (2018). *Reformed A level results: Do candidates who take the AS level achieve better grades?* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at www.cambridgeassessment.org.uk/Images/513414-reformed-a-level-results-do-candidates-who-take-the-as-level-achieve-better-grades-.pdf

Williamson, J., & Vitello, S. (2018). *AS level trends 2018*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at www.cambridgeassessment.org.uk/Images/519907-as-level-trends-2018.pdf

Further information on all journal papers and book chapters can be found on our website: www.cambridgeassessment.org.uk/our-research/all-published-resources/journal-papers-and-book-chapters/

Reports of research carried out by the Research Division for Cambridge Assessment and our exam boards, or externally funded research carried out for third parties, including the regulators in the UK and many ministries overseas, are also available from our website: www.cambridgeassessment.org.uk/our-research/all-published-resources/research-reports/

Statistics Reports and Data Bytes

The ***Statistics Reports Series*** provides statistical summaries of various aspects of the English examination system, such as trends in pupil uptake and attainment, qualifications choice, subject combinations and subject provision at school. The reports, mainly produced using national-level examination data, are available in both PDF and Microsoft® Excel format on our website: www.cambridgeassessment.org.uk/our-research/all-published-resources/statistical-reports/

The most recent additions to the series are:

- Carroll, M., & Gill, T. (2018). *Statistics Report Series No. 120: Provision of GCE A level subjects 2017*
- Gill, T. (2018). *Statistics Report Series No. 121: Uptake of GCSE subjects 2017*
- Gill, T. (2018). *Statistics Report Series No. 122: Uptake of GCE A level subjects 2017*
- Gill, T. (2018). *Statistics Report Series No. 123: Provision of GCSE subjects 2017.*

Data Bytes is a series of data graphics from Cambridge Assessment's Research Division, designed to bring the latest trends and research in educational assessment to a wide audience. Topics are often chosen to coincide with contemporary news or recent Cambridge Assessment research outputs. All *Data Bytes* can be found on our website: www.cambridgeassessment.org.uk/our-research/data-bytes

The following *Data Bytes* have been published since *Research Matters*, Issue 26 – those with interactive graphics are indicated with *(I)*:

- October 2018: *Tweets on A level results day*
- December 2018: *Parents' views on education around the world (I)*
- February 2019: *Geography assessment over time*
- March 2019: *Progression from GCSE (I).*

Sharing our research

Cambridge Assessment is home to the largest research capacity of its kind in Europe. In addition to our research publications, we provide a wealth of resources on our website including:

- *Insights* – a platform for sharing our views and research on the big education topics that impact assessment around the globe – at www.cambridgeassessment.org.uk/insights
- *Research Bytes* – short presentations and commentary based on recent conference presentations – at www.youtube.com/user/CambridgeAssessment1
- *#CamEdLive* – our online live debates – at www.youtube.com/user/CambridgeAssessment1
- Podcasts at www.youtube.com/user/CambridgeAssessment1
- Blogs at www.cambridgeassessment.org.uk/blogs

You can also learn more about our recent activities from Facebook, LinkedIn, Twitter and Instagram.

Finally, a reminder that all previous issues of *Research Matters* are available to download, in full and as PDFs of individuals articles, from www.cambridgeassessment.org.uk/research-matters

Cambridge Assessment

We have the largest research capability of its kind in Europe. It is this research strength that enables us to help teachers, learners and governments lead the way in education and unlock its power.

Our research is not just about ensuring our qualifications and services are the very best for learners, it's also designed to add to knowledge and understanding about assessment in education, both nationally and internationally. We also carry out research for governments and agencies around the world. It's all with one goal in mind – helping learners.

Our research is published in journals in the UK, around the world and on our website linked below. Here we publish research materials and resources, fact sheets, statistical reports and a regular series of graphics highlighting the latest research findings and trends in education and assessment:

cambridgeassessment.org.uk/our-research

Contents / *Issue 27* / Spring 2019

Cambridge Assessment
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

+44(0)1223 553985
researchprogrammes@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

3561869540

ISSN: 1755–6031