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	NNN	NNY	YNN	#Y#
	L	NNN	YNN	PNY	YY#
Expert compiled	H	YN-	PY-	NN-	##-
	K	N-N	Y-N	N-Y	Y-#
	M	-NN	-NN	-YY	-##
Computer compiled	J	NNN	YYY	NNN	Y##
	N	NNN	NNN	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.