Research and Analysis

Predicting predictability

Investigating question paper predictability and the factors that influence this through a question prediction exercise

Authorship

This report was written by Steve Holmes, Aneesa Khan, Nadir Zanini and Beth Black of the Strategy, Risk and Research Directorate.

Acknowledgements

The authors would like to thank the teachers and subject experts who participated in this research and provided us with the benefit of their insight and expertise. The authors are also grateful for feedback received from members of Ofqual's Research Advisory group.

Contents

Executive Summary

School leaders, teachers, researchers and other stakeholders have been interested in the predictability of tests for a long time, but it has been researched systematically on only a few occasions. It is important to recognise in the context of school examinations that the right amount of predictability is a good thing. There should be sufficient predictability to enable students and their teachers to have enough of an expectation of the nature of the demands and coverage of an examination to manage preparation and test anxiety. However, if there is too much predictability, then the expectations can lead to narrowing of preparation ('curriculum alignment') and perhaps even rote learning of responses.

Predictability has tended to be quite loosely defined by stakeholders and even researchers, leaving open 3 important questions:

- What are the factors that can lead to predictable tests?
- How can you measure predictability objectively?
- When does a test become too predictable?

The current study investigated these questions using the approach of collecting future question predictions and the rationales supporting these predictions from current teachers, and judging the similarity of the predicted questions to the actual questions.

The multi-phase study sampled papers from GCSE history, AS government and politics and A level psychology, with papers from two exam boards per subject. For each specification there were 3 phases:

1. Individual teachers made question predictions and supported these predictions with their reasoning

2. A subset of the teachers met to construct their best prediction for the questions on the next paper(s) in the series, for both the specification they currently taught and the other specification included for the same subject. These predicted papers were based on the predictions generated from the first phase as well as their own expertise

3. A review by subject experts of how accurate the predictions had been once the next live paper had been sat

The first and second phases generated rich qualitative data on the factors that teachers used to inform their predictions. Following analysis and coding of the rationales in phase 1, we devised a framework of factors that can narrow down future questions and thus define what can lead to predictability. This analysis included the relative frequencies of the factors for each specification, giving additional information on how the subjects and specifications varied. This was further expanded on in phase 2 when the reasoning behind the final question choices were also analysed and coded.

Although the precise factors most frequently identified varied across subjects and specifications, several common themes were identified, such as patterns of questions on past papers, the structure and wording of the content in the specification document, and the structure of the paper in narrowing down likely questions. The framework of factors should be borne in mind by test developers and item writers, as conforming to too many of the factors too frequently is likely to lead to predictable tests. Conversely, not following patterns suggested by the factors should introduce some desirable unpredictability in tests.

The final predicted papers produced in phase 2 were then compared to the live papers by subject experts in phase 3. This gave a measure of how much of the skills and knowledge required to answer the questions on each live paper were also necessary to answer the questions on the corresponding predicted paper. This overlap was generally around 40-50%. Government and politics papers were the most highly predicted, with one paper around 60%.

Interestingly, overall the predictions were slightly more accurate for the specification the teachers did not teach than the one they taught, so familiarity with the specification did not appear to help. These measures of correct predictions provide a first step in defining what may be considered to be acceptable and unacceptable predictability, although further work would need to be carried out to make this a practical approach to measuring paper predictability.

It must be remembered that the subjects and test papers included in this study were not representative of all subjects and test papers. For that reason, more factors may exist that were not captured here, and other subjects are likely to have different proportions of factors to those reported. However, this work provides a rich source of information that can help item writers avoid over-predictable tests.

Avoiding some of the factors detailed in this report could minimise some effects noted in a report on the sawtooth effect (Newton, 2020). This showed that over-predictable papers can contribute to undesirable and artificial improvements in student performance during the lifetime of a qualification specification and its assessments.

From a teaching perspective, predicting questions is an expected element of how teachers seek to prepare their students for exams. Some predictability is good - students should know what is expected of them and be supported in developing the skills needed to succeed in their exam. However, the research shows teachers do this to mixed effect and so it is always advisable that students are taught the whole curriculum – which will support their preparation for future study.

1. Introduction

The predictability of tests is of concern from a validity perspective. That is because overly predictable tests can lead to narrowed teaching - teaching to the test - and a less rich learning experience for students. Predictability is not a well-defined concept in educational assessment despite occasional concern over overly predictable tests (Byrne and Willis, 2004; Baird et al, 2016).

Overall assessment predictability is influenced by more than simply the predictability of questions that appear on papers and includes other factors such as knowledge of the assessment criteria and the alignment of textbook content with the questions (Baird et al, 2014b). However, predictable questions do perhaps give the most scope for a narrowed taught curriculum.

Overly predictable question papers allow teachers and learners to narrow their teaching and learning through having some degree of confidence in what they need to know, rather than having to learn the entire syllabus. Teachers can teach to the test and learners can restrict their exam preparation. This may go as far as having pre-prepared answers to predictable questions or topics.

A consequence of this would be that the test ceases to be a valid assessment of the syllabus, but rather becomes a test of a narrow set of knowledge and skills, and possibly a test of the quality of an individual's predictions and their recall. Under these circumstances the test may also cease to have any predictive validity of future performance.

However, it is important to recognise that a degree of predictability is desirable in all tests, as this may promote “test-wiseness”, an understanding of the requirements and expectations of the test. The format of the test, the structure of the papers and the style of questions asked should be familiar to candidates, as difficulties in accessing or understanding the questions would undermine candidates' ability to show what they know and undermine the validity of the test outcomes.

In addition, it is not usual for any one instance of an assessment to cover the entire syllabus – a sample of the whole syllabus is normally tested. However, there may be some areas of knowledge or skills that are considered to be so central to the subject that they should not be missed out in any instance of the test. Therefore, a degree of predictability may arise, although this can always be mitigated by varying the specific questions asked on these key topics.

What causes predictability, and how much test predictability may be appropriate has not been fully explored before. The current project uses teachers' predictions of future questions, and their justifications for these predictions to achieve two aims.

First, we wish to identify the sources of information that teachers consider when predicting questions. The framework of factors identified from the rationales for the teacher predictions can be used to inform the design and development of future and current tests. This should be of interest to test developers and item writers in avoiding overly predictable tests that allow questions to be predicted, and which, in the worst case, allow the taught curriculum to be narrowed.

Second, measurement of how accurate some of the predictions were for different tests may form the basis for a future method of measuring paper predictability in an objective way. Prediction accuracy may be seen as a proxy measure of how predictable individual papers are. This may help to define the range of acceptable predictability and in cases of tests falling outside this range, identify those requiring further scrutiny.

Both outcomes of the study will help in writing tests with an appropriate amount and type of predictability to make sure that candidates can show what they know while ensuring that they are taught the entire syllabus prior to the test. Although this work is focused on the predictions of questions appearing on papers, the framework considers factors beyond the question papers. We recognise the importance of other broader factors on overall q5ualification predictability.

1.1. Previous research

Several pieces of work have directly assessed qualification predictability, usually from the perspective of a regulator or exam board. It is not immediately obvious how to define and measure predictability. This research has taken a largely qualitative approach, with subject and assessment experts identifying issues, or sometimes using rating scales to try and quantify aspects of predictability. Although these analyses have touched upon the factors that lead to predictable questions they have not been systematically collected.

In addition, quantitatively measuring and then defining the optimum amount of predictability to build into a test has not been attempted previously, although it was suggested as an approach by Murphy et al (2012), as described in section 1.1.2.

The 3 key studies looking into the predictability of question papers are detailed below.

1.1.1. Ofqual predictability study, 2008

The first study to attempt to measure predictability was carried out by Ofqual (2008). Qualifications were selected for the main study through a review of materials by experts and a search for specifications where concerns over predictability had been raised in the regulator’s monitoring reports. Subject assessment experts then reviewed question papers, mark schemes, examiner reports and specifications for the 10 selected subjects, covering both GCSE and GCE level. The experts also considered additional materials such as textbooks and exam board supporting materials.

For each subject, two or three experts rated and commented on the materials on five factors on a scale from 1 to 11. The factors were quite broad, and included the relationship between the specification content and the question paper, and nature and format of questions within and between series. On completion of the rating exercise the experts were also asked for their predictions for the next paper in the series. The predictions required were not constrained and took the form of free text comments, rather than requiring specific question predictions.

For those qualifications identified as having potentially problematic predictability, interviews with teachers and students were also undertaken. This was to see how the experience of taking these qualifications might reflect the expert findings. For some assessments limited analysis of item functioning took place to determine whether items judged to be predictable showed less spread in the marks or higher average marks than might be expected from their judged difficulty, or were disproportionately popular where they were optional questions.

Finally, a review of scripts took place, to see whether the features identified by the experts in the assessments were having the expected effect on candidate work. This study looked at a limited sample of assessments that had been selected because there were already some concerns about them, and so although this study identified some potential problems of predictability in some papers, predictability was not thought to be a major issue across all GCSE and GCE assessments.

The main risks to over-predictable assessments came from 3 areas. The paper structure could allow candidates to narrow their coverage of the content, often through choice of optional questions, with little risk. The syllabus content was often over-specified, and the question setters often followed the wording of the content very closely so that questions were well aligned with the specification. However, this narrowed the range of questions they could ask so they became formulaic. There were also potential issues around the cycling of topics across papers being too predictable.

The experts did note that the predictable nature of the question stems and command words led to helpful and positive predictability which allowed the candidates to clearly understand what was required in the examination. In contrast, the experts identified unpredictable assessments with unexpected question types as just as much of a risk to validity as over-predictable ones. Both factors could lead to a test where test-taking skill and preparation would be more important than knowledge and understanding.

This study was described as being somewhat pilot-like in nature, and it is clear that the forms used to capture expert judgements were a little imprecise. The report states that different experts sometimes interpreted what the forms were asking for differently. Certain parts of the data collected, such as the predictions for the next paper in the series, may not have been that useful given that they were not referred to in the report in any detail. However, the independent work provided a good starting point for the group discussions that followed, which allowed consensus judgements about each specification to be reached and followed up in interviews and script scrutiny.

1.1.2. An outline for a predictability study – Murphy et al

Murphy et al (2012) produced an unpublished report for the awarding body Pearson detailing a full outline for a planned future predictability study, building on the Ofqual study and extending the methodologies used. Despite no published work resulting from these suggestions, it provides a useful summary of a variety of different approaches that could be adopted. Similar to the Ofqual study, Murphy et al noted that predictability is a larger issue than simply question writing. They classified problems in 4 categories:

- *Policy and regulation*. The need to ensure comparability across exam boards and over time has led to constraints on exam boards. Common content, assessment objectives and similar assessment types can all increase predictability.

- *Awarding body processes and cultures*. The need to meet the regulations can lead to very precise specification of all details of the assessments, to show compliance. Sample assessment materials produced as part of the specification can also significantly constrain the assessments that follow.
- *Examining and resources*. Schools are provided with a variety of additional materials and information about exams. Training and materials provided to schools, examiner reports detailing how candidates performed in the exam can all lead to a focus on exam preparation which may become undesirable.
- *School-level responses*. Schools and teachers will take different views on how much they need to teach to the test and focus on exam technique. In the extreme this may be undesirable.

Research could be focused at all or any of the levels. They again noted that predictability is relative, where the optimal amount of predictability may vary across subjects, and therefore there is always a need to include inter-board comparisons when evaluating predictability within a subject area.

They laid out a similar outline plan to the Ofqual study, with small groups of subject experts reviewing the full suite of assessment materials, including examiner reports, specifications and, in addition, the sample assessments. Similarly, following this review stage, using slightly more precisely defined categories, but simpler rating scales than the Ofqual study, candidate scripts and item level data could be analysed to verify whether factors identified in the review were evident in responses.

They also intended to follow up this desk-based work with centre visits to interview teachers and students to see whether classroom behaviours reflected the reviews. Interviews with senior examiners were also suggested in order to draw out any longer-term influences on, and trends in, predictability. They placed more emphasis on the use of future question prediction as a way to check the underlying paper predictability than was evident in the Ofqual study. They also noted the potential for examiners during the live marking to note down evidence of predictable questions as they marked, particularly where they were marking an item one at a time online rather than whole scripts. Other, wider ranging approaches to consider, such as specification development and classroom-based studies, were mentioned.

1.1.3. Irish leaving certificate – Baird et al

A few years later Oxford University Centre for Educational Assessment and Queen’s University, Belfast were commissioned by the Irish State Examinations Commission to carry out a substantial investigation of predictability in the Irish Higher Level Leaving Certificate following several years of media concern over these exams (Baird et al 2014a,b, 2016).

This study covered 6 subjects and used a variety of approaches, including a similar approach to expert rating of examination materials to the Ofqual (2008) study, interviews and surveys with teachers and students and data analysis relating student outcomes to their survey responses for 3 subjects in the Irish Leaving Certificate.

The subject expert reviews, by two or three experts per subject, included checking the syllabus, teachers’ guide, question papers and mark schemes, and student scripts. The task was focused on comparing consecutive years of the assessments and rating their similarity on 8 dimensions. These were intended difficulty, content, skills, command words, question wording, question type/format, layout, and resources. They used a scale from 1 to 4. Following the independent rating, workshops were held to agree the consensus view through discussion. These reviews suggested that some of the assessments had some problematic predictability that might promote superficial or rote learning, while other assessments were fine.

A related earlier piece of work (SEC, 2012, cited in Baird et al, 2014b) had also suggested that topic choice in some Leaving Certificate subject examinations should also be reviewed. Surface features such as question type/format and wording and paper layout were rated as predictable, which was viewed as a positive thing as it made the papers accessible to candidates.

The interviews with teachers and students and the student survey were useful in elaborating what the effects of any predictability in the assessments were on the learning outcomes for students. Was teaching narrowed by focusing on the test and the topics or skills likely to be tested? What effect did the students' perceptions have on their test preparation and were outcomes related to their perceptions?

Teachers did not have any major concerns about predictability, and felt that the topics coming up were not predictable. Students did sometimes report predicting topics. But an interesting finding was that those students who believed that the curriculum could be narrowed in their exam preparation performed worse on the tests, even when socio-economic status was controlled for. Yet there was unfortunately no control for the key measure of prior attainment.

Overall, some predictability was identified in some subjects. This related to the frequent appearance of some topics, formulaic question styles, narrow coverage of the syllabus or the availability of optional routes, which meant that not all the syllabus needed to be learnt. There was a strong perception that many papers were rewarding knowledge recall instead of higher order skills, and that this, together with somewhat predictable questions, provided scope for pre-prepared answers to be used.

1.1.4. Other research

There is an extensive literature on the effects of high stakes testing on teaching (washback) which was recently reviewed in Baird et al (2014a) and so is not repeated here. This literature frequently relates to the effect of perceived or real predictability on narrowing teaching and learning, but does not attempt to measure or assess predictability in specific assessments. Although not directly investigating predictability, other papers have touched on wider issues arising from the predictability of tests.

For example, Crisp et al (2008) considered the effect of candidates' expectations of questions on their performance. This work was based on the idea that candidates have recognised or been taught the predictability in the tests, in terms of the way questions are usually asked, and the alignment of question types to topics. These expectations mean that questions which ask something unexpected, either in the way they are phrased or the actual response that they require, may sometimes prove problematic to the candidate, particularly where the wording is suboptimal. This can occur when the first part of the question activates an existing schema which overrides the detail of later parts of the question, particularly in a stressful examination situation.

Crisp et al modified questions to see how question wording or layout lead to misunderstandings of the task, and to see how this could be avoided in writing future questions. They suggest that in some instances "test-wiseness" can be unhelpful when slightly unusual (but valid) questions are asked. These kind of issues, while not directly related to how predictable individual tests are, arise because of the regular patterns in questions and topics that occur.

Daly et al (2012) considered changes to A level assessments to add stretch and challenge in 2008. The intention was for new assessment styles to encourage better teaching with test preparation being less narrowly focused. In interviewing teachers and students prior to the first sitting of the new assessments, they found that there was indeed a substantial focus on the content of the exams, which was influencing the teaching and learning. There were concerns from teachers over how they would adapt teaching to meet the new assessments, which had greater content coverage in individual tests. Teachers felt this was putting additional pressure on them, indicating that the focus of teaching was indeed often on the tests rather than the syllabus, which had not been increased.

This suggested that from the point of view of the reform, the added stretch and challenge in the assessments would have significant effects on teaching. However, it was also indicative of potentially too much focus on tests over and above broader learning, and that situations where there was possibly too much transparency around the requirements of the assessments and the marking criteria could potentially lead to undesirable washback into teaching. This can only really occur in a situation where papers are sufficiently predictable to allow a narrowing of the taught curriculum.

Finally, it is worth noting recent attention on the social and cognitive processes of writing questions. Johnson, Constantinou and Crisp (2017) considered the process of writing individual questions and devised a model of this process, which Johnson and Rushton (2019) extended to the construction of whole papers. Predictability was not the focus of this research, but would clearly be a factor when considering the appropriateness of questions when the test forms one of a series over time for the qualification. The current work may feed into the thinking of such item writers carrying out these kind of exercises.

1.2. Current study

When considering the predictability of test materials, the studies detailed above have taken a wider view of predictability. This included considering aspects of assessments beyond the questions themselves that could seem problematic to subject experts. The data obtained has not always been robust enough to draw firm quantitative comparisons between specifications or question papers, being based on small samples of experts who may have used the rating scales in slightly different ways.

In this study we restrict ourselves to the prediction of individual questions, leading on to the prediction of whole papers, as an extension of one of the approaches suggested by Murphy et al (2012). By getting a substantial number of experienced teachers to make predictions and justify their choices, we can collect rich data on all of the factors they use to make their predictions, potentially including factors that we might not have considered. This differs from previous approaches which have largely provided categories of predictability to be evaluated, with some additional comments recorded.

It is worth noting that while each individual teacher may not take into consideration every factor, across a sizable group of teachers we are likely to get good coverage of most or all possible factors likely to influence predictions on these tests. We will also have enough predictions to gather some quantitative data on the relative importance of these factors. This will help to identify factors which may be problematic in increasing predictability in the tests.

There were three phases to this study:

1. For 6 specifications (2 specifications in each of GCSE history, AS government and politics and A level psychology), approximately 10 teachers who taught each specification made independent predictions of the questions that they thought would appear on 1, or sometimes 2, paper(s) in the next sitting of the series in summer 2017, based on reviewing the specification content and the most recent papers in the series. They also stated the factors that influenced these predictions and their reasons for making the prediction.

2. At a day-long meeting for each of the specifications, 4 of the teachers who had made the phase 1 predictions were provided with all of the phase 1 predictions, including their own, and worked together to narrow these down to one set of predictions – a final predicted future paper. They also repeated this process for the other specification in their subject which they did not teach, creating ‘home’ and ‘away’ predictions. These meetings were recorded to capture the reasons for the final choices.

3. Following the sitting of the summer 2017 tests, the predicted papers were evaluated by subject experts for how close they were to the questions that appeared on the live summer papers. This allowed us to measure how much of the skills and knowledge assessed in the summer 2017 test was accurately predicted. We treat this prediction accuracy as a proxy measure for how predictable each paper is.

Using this indirect measure of the paper predictability, the relative accuracy of the predictions can be compared. Although this cannot definitively tell us what an appropriate level of predictability is, such comparisons can highlight assessments at the extremes of predictability for further review.

In summary, this study is designed to provide results that can help to answer the following research questions:

1. What are the factors that can lead to test predictability?
2. How can we measure predictability, and how predictable are a sample of current tests?
3. When does a test become too predictable?

2. Methods

2.1. Choice of papers

Papers were chosen from 6 GCSE, AS and A level specifications - 2 specifications from each of 3 subjects. The subjects were chosen based upon the structure of the papers, any historical concerns over predictability in the assessments, and coverage of both non-reformed and recently reformed specifications.

AS government and politics and GCSE history were both going through a process of reform, with the summer 2017 paper representing the last sitting of the current specifications, and therefore these subjects should be near to maximum predictability. A level psychology was a newly reformed subject, with no live papers sat, and only sample and practice papers available, and on that basis might be seen as having low predictability. The specifications within each subject were chosen based on entry size and also a balance of specifications across the 3 exam boards included in the study.

Within each specification, papers were chosen that represented approximately the same amount of testing time and a similar number of questions, particularly between the specifications within a subject. This meant that for 3 specifications one paper was used, while for the other 3, two (shorter) papers were used, giving 9 papers in total.

Papers were also chosen which had the maximum amount of overlap in content between the two specifications within a subject. This was because we wanted teachers of particular specifications to be reasonably familiar with the content for the other specification in their subject. This was important for the design of the meetings we ran to decide the final predicted papers where ‘away’ predictions were made (see Section 2.3).

Where there were optional units within a specification, as was the case with the GCSE history specifications, popular units with large numbers of candidates sitting them were chosen to make it easier to recruit teachers to make predictions. Table 1 lists the papers included in this study.

The papers varied in their structure and Appendix A details each paper in turn. In summary, some papers had quite fixed structures, with a set number of questions, patterns of question tariff and usually the same or similar question stems or types on every iteration of the paper. For these fixed-structure papers, content was either assessed with largely the same question structure in each series, as occurred for government and politics, or the content was rotated around the sections of the paper, as was the case with history.

The psychology papers by contrast were less rigidly structured, with more variation in number of questions, question tariffs and stems within each section of the papers. However, the specification content to be sampled was mostly allocated to specific paper sections.

Table 1: *Summary of papers included in the prediction exercise.*

Subject / Specification	Level	Duration (mins)	Number of questions
Government and politics	AS		
Specification 1		90 90	12 12
Specification 2		80 80	12 8
History	GCSE		
Specification 1		120	25
Specification 2		75 75	7 8
Psychology	A level (reformed)		
Specification 1		120	Flexible (approx. 20)
Specification 2		120	Flexible (approx. 22)

2.2. Phase 1 - Independent predictions by teachers

2.2.1. Materials

To help carry out the exercise, our participants were sent copies of the specification document, which contained the syllabus content, and a set of past papers. The 4 specifications for government and politics and history had run largely unchanged for several years, and there were 4 years of past papers available from the exam board websites. We provided our participants with the four summer series papers from 2013-16. For the new psychology specifications, no live papers had been sat, so participants were asked to consider the sample assessment and any additional practice/sample papers available: two papers for specification 1 and three for specification 2.

In addition to the past papers and specification documents, the teachers were asked to complete a report/template. These response documents were bespoke for each specification, to be completed electronically by the teachers. They contained an outline of the task, followed by specific information on how the document was structured and precisely what needed to be entered into it, followed by the tables to be filled out.

Appendix B includes a copy of the outline of the task we included in the response documents. This outline was the same for each specification. The intention was to encourage the teachers to think carefully about their predictions and to consider a wide variety of factors that could influence them.

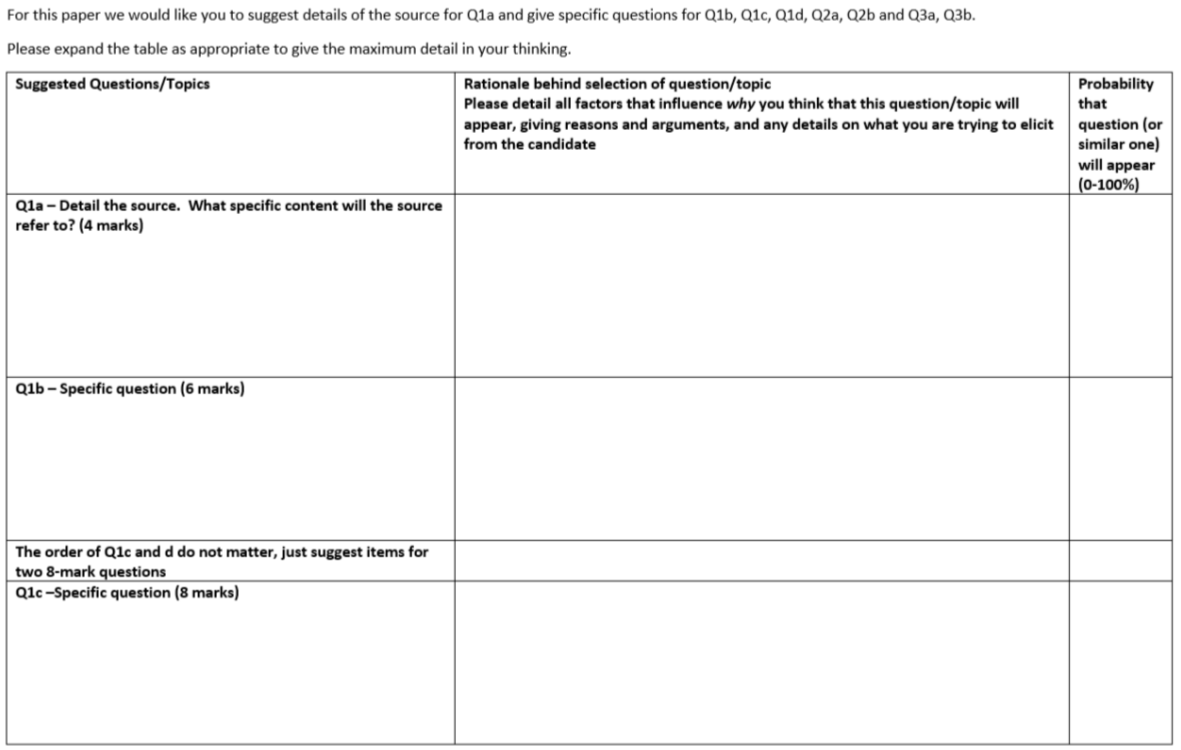
The outline was followed by a short section tailored to each specification detailing what needed to be completed in the tables that followed, and how the tables were structured, together with details on returning the document to us. This section also encouraged our participants to make more than one prediction per question slot on the paper, where they wanted to. To help with making multiple predictions the table that followed contained a column in which to detail confidence in the prediction so we could see which predictions were thought most likely to occur.

The purpose of this stage was to elicit factors that influenced their predictions. It was also to provide predictions for the phase 2 meetings to consider so we were not concerned that making multiple predictions could make it more likely that these included what turned out to be ‘correct’ predictions.

The rest of the response document was filled with tables which the teachers needed to fill out. They were designed to collect the most appropriate type of prediction for each question or sub-section on the paper. For example, some questions had a fixed wording or format, and only the stimulus materials changed. In this instance it was not useful to ask for the question wording, rather a prediction of the material itself was most useful. Some papers had a fixed format, with repeated patterns of question types and tariffs, where the content assessed for each question varied. Other papers had variable tariffs and numbers of questions within fixed sub-sections.

So sometimes we requested predictions for specific numbered questions, while other times we asked for a set of predicted questions and tariffs that would add up to make a whole sub-section. Figure 1 shows two examples from the response documents, showing (a) a history paper with a fixed structure and (b) a less structured psychology paper.

a)

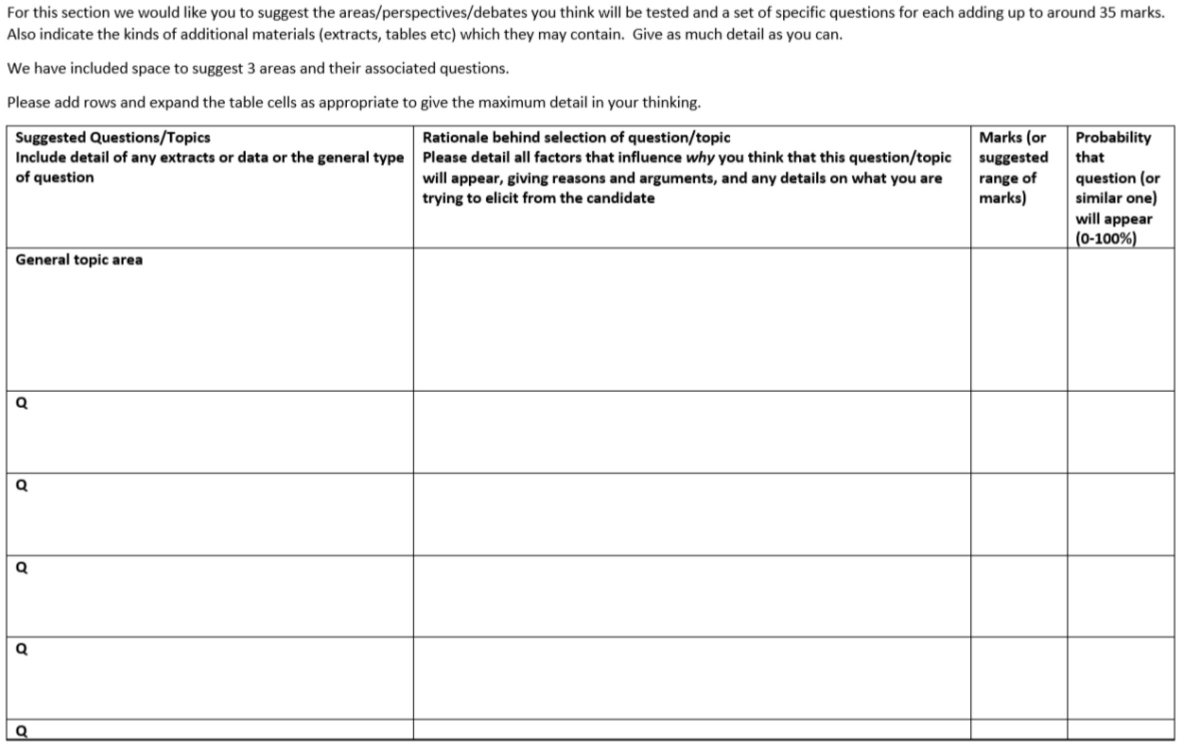


For this paper we would like you to suggest details of the source for Q1a and give specific questions for Q1b, Q1c, Q1d, Q2a, Q2b and Q3a, Q3b.

Please expand the table as appropriate to give the maximum detail in your thinking.

Suggested Questions/Topics	**Rationale behind selection of question/topic** **Please detail all factors that influence *why* you think that this question/topic will appear, giving reasons and arguments, and any details on what you are trying to elicit from the candidate**	**Probability that question (or similar one) will appear (0-100%)**
Q1a – Detail the source. What specific content will the source refer to? (4 marks)		
Q1b – Specific question (6 marks)		
The order of Q1c and d do not matter, just suggest items for two 8-mark questions		
Q1c –Specific question (8 marks)		

b)



For this section we would like you to suggest the areas/perspectives/debates you think will be tested and a set of specific questions for each adding up to around 35 marks. Also indicate the kinds of additional materials (extracts, tables etc) which they may contain. Give as much detail as you can.

We have included space to suggest 3 areas and their associated questions.

Please add rows and expand the table cells as appropriate to give the maximum detail in your thinking.

Suggested Questions/Topics **Include detail of any extracts or data or the general type of question**	**Rationale behind selection of question/topic** **Please detail all factors that influence *why* you think that this question/topic will appear, giving reasons and arguments, and any details on what you are trying to elicit from the candidate**	**Marks (or suggested range of marks)**	**Probability that question (or similar one) will appear (0-100%)**
General topic area			
Q			
Q			
Q			
Q			
Q			

Figure 1: *Examples of the tables that teachers were asked to fill out to record their predictions. The two images show tables for a) a fixed structure paper (history) and b) a more flexible structure paper (psychology).*

2.2.2. Participants

For each specification, we aimed to recruit 10 teachers to make independent predictions. Letters were sent to the head of department at a sample of centres we identified as having entered candidates for that specification, or the companion AS specification in the case of the new psychology specifications.

We invited teachers who currently taught that specification and who considered that they had a good awareness of the questions that appear on these exam papers to contact us. Only one teacher per centre was recruited, to ensure the predictions were independent, and teachers were recruited on a first come first served basis with no selection criteria.

Table 2 lists the number of teachers within each specification who took part. Due to withdrawals the number completing the task was sometimes fewer than the target number. All participants were paid for their time.

Table 2: *Number of participants for each specification*

Subject / Specification	Number of participants
Government and politics	
Specification 1	7
Specification 2	9
History	
Specification 1	9
Specification 2	10
Psychology	
Specification 1	9
Specification 2	11

After completing the first phase of this study the teachers completed a short survey which included questions on their experience which we report here. Table 3 shows that this was quite an experienced group of teachers, with over 10 years of teaching experience on average. They had also been teaching specifications from the same exam board for a considerable time, suggesting that there had not been a great deal of switching between exam boards for these teachers. For our history sample, there appears to be a little more switching.

Table 3: *Teaching experience of the participants in this study*

Subject	Mean years teaching subject (range)	Mean years teaching specification(s) from current exam board (range)
Government and politics (n=16)	14.1 (1*-32)	11.6 (2-28)
History (n=19)	13.8 (3-37)	8.0 (2-21)
Psychology (n=20)	10.5 (3-20)	9.8 (1-20)

* One teacher had been teaching for 10 years but had only started teaching government and politics a year ago

2.2.3. Procedure

We sent the teachers electronic copies of the specification, past or sample or practice papers, and the response document to be completed and emailed back to us. They had 2 weeks to make their predictions and return the document to us. Although we only provided past papers over the 4 most recent summer series, we did not discourage the teachers from going further back if they had copies of older papers and wished to make use of them. We also did not deter them from using other sources of information to make their predictions.

The introduction to the response document made it clear that the teachers need not worry about the precise wording of the questions they predicted as it was the general sense of the question and the content and skills it assessed that was important. They were also prompted to give as many predictions as they wanted for each slot on the paper, not to necessarily limit themselves to one.

Upon return of the response document, we sent a short survey to the teachers with questions regarding their choice of specification, and their previous experience of predicting topics and questions.

2.3. Phase 2 - Meetings to produce final predicted papers

2.3.1. Materials

Printed copies of simplified versions of the response documents from the first phase were provided at the meetings, together with copies of the specification documents and past papers. The rationale column was removed from the response documents, and the independent predictions were collated by question slot/paper sub-section. The rationales were removed due to time constraints in the meetings – there would not have been time to read and discuss this reasoning – but also to avoid particularly elaborate or persuasive rationales from influencing the participants.

2.3.2. Participants

Four of the teachers for each specification who had made predictions in phase 1 attended the meeting[1]. As soon as we received an expression of interest to take part in phase 1 we also asked for their availability to also attend a day-long meeting. Places were allocated based on the order we received their confirmation of interest, with no selection criteria. This recruitment took place before they carried out their independent predictions. They were paid for both phases of the study.

2.3.3. Procedure

In all meetings the teachers first worked on constructing the predicted paper for the specification they currently taught, which we call the ‘home’ papers, then worked on the other specification, which we call the ‘away’ papers. Within a paper the teachers chose how to work through the papers, especially which questions or sections to start with. Sometimes this involved returning to earlier predicted questions and adjusting them to create a coherent whole paper.

An Ofqual researcher was there to record the final choices, prompt the teachers for their reasoning and to keep to schedule. On rare occasions where the teachers were unable to come to a decision on question wording, or there was a split decision between two questions, they would make a decision to break the deadlock.

All the meetings were audio recorded to save the discussion for later analysis. All participants consented to this recording. The output from each meeting was predicted papers comprising the correct number of questions and marks, for each of the home and away specifications. Because there were 2 teams of teachers working within each subject, for each specification there were 2 sets of predicted papers, one created by teachers who taught the specification and one from teachers who did not.

2.4. Phase 3 - Subject expert evaluation of predictions

2.4.1. Materials

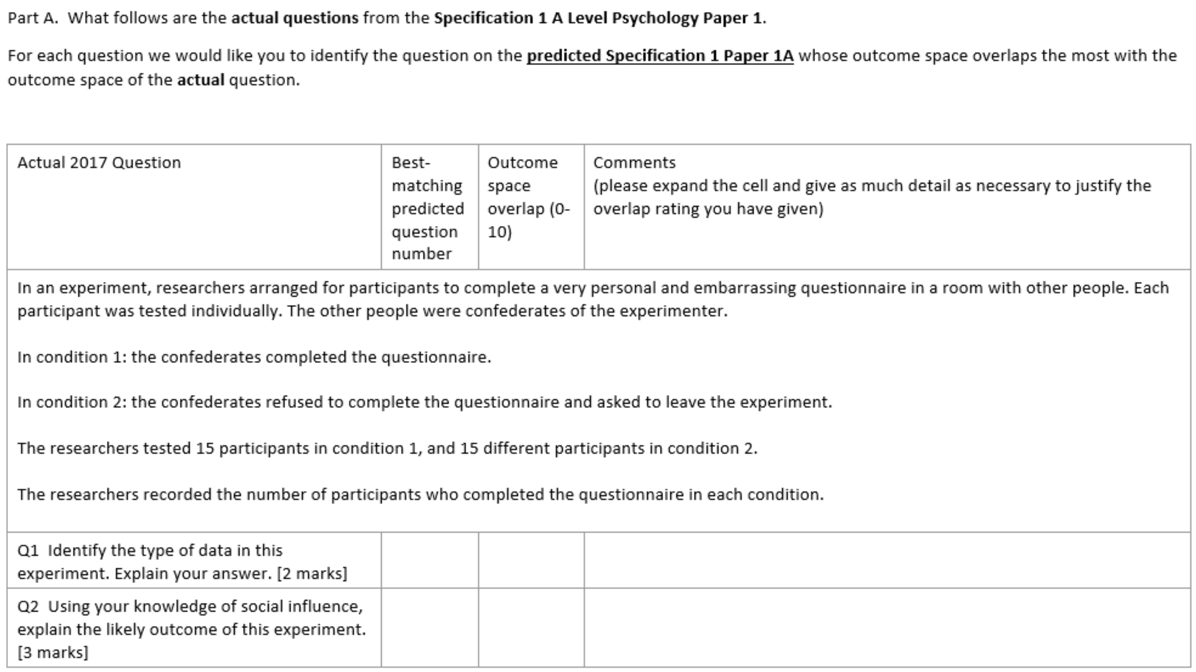
We transcribed the predicted papers into PDF documents with the questions structured as per the live papers. We did not indicate the source of the papers, just identified them as predicted paper A, predicted paper B and so on. Alongside copies of the live summer 2017 papers sat by candidates, we provided our expert raters with a response document in which to record their ratings. This document also contained the description of the task, and how to complete it. Appendix C includes an example of the outline of the task we included in the response documents. This information was modified only to account for the different number of predicted and live papers per subject.

In addition to this information, at the end of this document we also attached a 2 page description of outcome space, partially adapted from Pollitt and Ahmed (2008).

[1] In one instance a teacher had to withdraw at short notice. There was not time to replace them so only 3 teachers took part in the meeting.

Outcome space is the range of possible responses that students may produce to show that they have the required skills and knowledge which a question is intended to assess. We wanted our experts to understand and use this concept in deciding on the degree of knowledge and skills overlap between questions. Thinking about this outcome space for the questions on the live papers, and then thinking about how much of this space the predicted questions might overlap on should allow the experts to identify the best-matching question, and give an estimate of how much overlap there was. See Appendix C and Appendix D for the more detailed description of outcome space which we provided.

The rest of the document contained tables as shown in Figure 2. This has the questions from the live summer 2017 papers listed in the first column and columns for the subject experts to fill out detailing which question on the predicted paper was the closest match to the live question, and why. The instructions made it clear that the same predicted question could be used as best match to more than one live question. The document was designed to give a measure of overlap for all questions on the summer 2017 papers, so that it was possible to determine how much of the live paper had been partially or fully predicted.



Part A. What follows are the **actual questions** from the **Specification 1 A Level Psychology Paper 1**.

For each question we would like you to identify the question on the **predicted Specification 1 Paper 1A** whose outcome space overlaps the most with the outcome space of the **actual** question.

Actual 2017 Question	Best-matching predicted question number	Outcome space overlap (0-10)	Comments (please expand the cell and give as much detail as necessary to justify the overlap rating you have given)
In an experiment, researchers arranged for participants to complete a very personal and embarrassing questionnaire in a room with other people. Each participant was tested individually. The other people were confederates of the experimenter. In condition 1: the confederates completed the questionnaire. In condition 2: the confederates refused to complete the questionnaire and asked to leave the experiment. The researchers tested 15 participants in condition 1, and 15 different participants in condition 2. The researchers recorded the number of participants who completed the questionnaire in each condition.			
Q1 Identify the type of data in this experiment. Explain your answer. [2 marks]			
Q2 Using your knowledge of social influence, explain the likely outcome of this experiment. [3 marks]			

Figure 2: *Examples of the response document showing 2 questions. Subject experts filled out columns 2-4.*

2.4.2. Participants

We recruited 3 independent subject experts for history and psychology, and 2 experts for government and politics to carry out independent comparisons of the final predicted papers to the live summer 2017 papers. The experts were paid for their time.

2.4.3. Procedure

The subject experts were sent copies of the predicted papers and the response document by email and asked to return their ratings within 4 weeks. As mentioned in the materials section, they were instructed to use the outcome space idea of Pollitt and Ahmed (2008) to decide how much the knowledge and skills assessed by the questions on the summer paper(s) overlapped with questions on the ‘home’ and ‘away’ predicted papers.

We did not specify exactly how the experts should make their decision on outcome space overlap; the guidance provided is included in Appendix D. Broadly speaking, they were to imagine that a candidate knew only what had been assessed on the predicted paper, and nothing else. But they knew this to a level that would achieve full marks on the predicted paper. For each question that came up on the live paper they were to estimate how much of the skills and knowledge required to fully answer the question this imaginary candidate would have.

Zero and 10 were clearly defined, but we made it clear that a score of 5, indicating that the candidate would have half the knowledge and skills required to fully answer the question, would not necessarily map to half marks. The expectation was that the precise content chosen would carry more weight in this overlap than the kind of cognitive operation (skill) required, given that without knowledge of the content on the live paper, knowing how to do something would not get you that far in answering the question.

3. Results

3.1. Phase 1 - Analysis and coding of individual predictions

The returned response documents from phase 1 were used to devise a framework of factors which can influence predictions, and to produce counts of how frequently these factors were mentioned. Three Ofqual researchers took part in devising the framework and carrying out initial coding. The framework was built up in an iterative process as follows:

1. We initially started with the factors listed in the response document (see Appendix B). Independent analysis of randomly-chosen response documents spread across the specifications was undertaken, with each coder independently making modifications to factors or adding new factors as they came across rationales that did not fit well into existing factors.
2. Meetings were then held to discuss and refine the suggested changes and agree a common modified set of factors.
3. More practice coding then took place followed by a further meeting at which the final set of factors were discussed in detail to clarify the meaning of them all, including strict inclusion or exclusion criteria.
4. Final coding was then carried out by 2 of the researchers, due to the unavailability of the third, with each teacher document coded by both.
5. As a final check early in this coding, both researchers initially coded the same small set of response documents to test the framework for completeness and usability, followed by meetings to compare coding and ensure alignment. However, even when looking at the teacher documents together, there were still sometimes disagreements about whether a particular comment should be coded against particular categories, and some inconsistencies were inevitable.

For history and government and politics, each question slot on the papers was coded against each factor as ‘0’ (factor not mentioned) or ‘1’ (factor mentioned). Inter-rater coding agreement was calculated from the number of cases that a question was coded ‘1’ against a factor by both researchers compared to the number of cases where either or both researchers coded the response as ‘1’.

This gave the proportion of times that both coders agreed out of the total cases where they either agreed or disagreed. Because the coding grid was very sparse, cases where both researchers coded the factor as ‘0’ were ignored. That is because there would have been artificially high agreement between coders if all these cases with both coding ‘0’ were included. This sparseness also meant that no adjustment for chance agreement, such as Cohen’s Kappa (Cohen, 1960) was made.

Overall agreement as shown in Table 4 was reasonable considering the unconstrained, and sometimes ambiguous, nature of the justifications made by the teachers, and the slightly overlapping nature of some of the categories. For this, see the next section.

Table 4: *Inter-rater agreement for each specification of government and politics and history combined across all teacher predictions*

Subject / specification	Percentage agreement (both coders indicate factor was present)
Government and politics	
Specification 1	59%
Specification 2	62%
History	
Specification 1	60%
Specification 2	64%

Psychology was coded slightly differently. Because of the flexible structure of the papers the response documents were not divided into individual question slots, but contained a variable number of predicted questions per teacher for each section of the paper. Therefore our coding was done at the section level, with counts of how many suggested questions in the section had rationales within each factor. The coding was therefore not restricted to '0' or '1'.

Perfect agreement of the counts in each section between coders would tend to be low, as any differences between coding individual questions would be multiplied through the combination of predicted questions within each section. So, we do not report this agreement, but note that the same process was taken in coding and so inter-rater agreement for individual questions in psychology will be similar to the other subjects.

3.2. Phase 1 - Factors influencing teacher predictions

3.2.1. Identified factors

Each factor identified in the coding task is detailed in this section, grouped into over-arching themes, with a brief summary of the kind of observations it includes and sometimes how its occurrence varies across subjects. Some factors are more specific than others, leading directly to particular predictions, while others are broader, and provide reasons for making particular predictions more likely. In some instances factors are counted as sub-categories of other factors. In these instances, the intention was to code only against the more specific category, not both.

3.2.1.1. Factors related to appearance of questions/topics on past papers

1. High frequency of topic appearance makes it likely (it usually comes up)

This question or topic comes up often in some form on past papers. This may be because it is an important part of the specification, or it may be because fixed question types limit topic choice for some questions, so it does occur frequently – it is a ‘popular’ question. This factor does not need any reference to cycles or lack of appearance recently, although these may occur as well.

In practice, the comments are usually quite straightforward observations about the regularity of the question appearing, or its ‘popularity’ with the examiners and/or students. There is overlap with factor 3 below, which is a sub-category of this factor in which the topic appears a lot, but in different forms, so they are effectively assessing different skills. Therefore comments in this category are frequently simple ‘often comes up’ type comments.

2. High frequency of question type appearance makes it likely (it usually comes up)

This factor refers to frequent appearances of the particular type of question (not topic) making it highly likely to appear. Factors 1 and 2 are spilt across subjects. In history and government and politics, where the paper has a fixed tariff and largely fixed question stem structure, this factor is not usually relevant, but for psychology, which has an open structure but fairly fixed coverage of topics, observations tend to be more about types of questions being used frequently.

3. Topic/question type has come up frequently, but in a different form / place

Comments here refer to something being a popular topic as above, but there will be explicit mention that it has come up in different forms (tariffs, command words) on past papers. Often comments start with the topic always coming up, often also related to comments around centrality to specification, or amount of teaching time spent on it, but that it has moved around and therefore this year it is likely to appear in a particular slot on the paper.

There will often be a process of exclusion where the past forms it has appeared in are discounted and so the prediction becomes more likely. The prediction could incorporate a variation in the wording whilst largely asking a similar question, but usually involves a different analysis for the same topic area. Comments here may also be seen in association with topic cycling comments. In the psychology predictions this may relate to the rotation of question types across sections of the paper or topics.

4. Past patterns of question type/topic cycling lead to this topic

This factor includes references, maybe only implied, to patterns, sometimes a fixed cycle, of appearance. This observation may be explicit, or it may be implicit from an analysis of what has occurred on the past papers with a logical conclusion as to what is due or likely to follow. This factor is more specific than non-appearance last year (factor 5 below) and it may be considered a sub-category of this factor since non-appearance is implied by it.

The related issue of old specifications that have assessed everything needing to recycle old questions also fits here. This may also include the cycling of types of questions across different sections in papers without a fixed question structure, as is the case with psychology. In practice, comments rarely refer to an absolutely fixed pattern, as there is usually some randomness/unpredictability in the cycle of topic appearances.

5. Non-appearance last year/recent years/ever increases chance of appearance

This factor records points about ‘the topic being overdue’ as it has not appeared for a certain time, or in the sample assessments for psychology. This category may include ‘never been used’ or ‘not seen since <year>’. Generally, this is a broad category, with much less specific analysis of patterns than factor 4. Those that are more analytical will usually be categorised within factor 4.

3.2.1.2. Factors relating to the content in the specification document

6. Importance/centrality in specification makes it likely it will come up

This topic is thought to be so important within the specification that it almost always comes up. The observations are that it has to be assessed each year in some way. This factor is related to factor 1, but the comments refer to the specification, not occurrence of topics or questions on the papers themselves. Sometimes these comments will refer to the way that the specification structures the content, with a predicted question assessing indicated ‘key topics’ or ‘key questions’ in the specification.

7. Alignment of wording to specification content

These comments refer to the wording of the content being reused almost verbatim in the wording of questions. There is some overlap with factor 6 since some of the predictions will be drawn directly from key topics or key questions listed in the specification, but they are not always question predictions for particularly important pieces of content. They will always contain an explicit reference to the wording in the specification. This factor is about narrowing down the question wording using the specification content when the topic area of a prediction has been decided on.

3.2.1.3. Factors related to the appropriateness of topic for the type of question

8. Topic/question type fits position on paper (size, type – event/treaty etc)

This factor is about the alignment of the chosen topic with the type of question being asked. The topic area may be of the right size or type to fit a particular tariff, or fixed question stem. For example, there may be observations that a question that always starts ‘explain the causes…’ only works with some topics, or that ‘this question is always about a …...’ to narrow down the choice of topic.

There are also comments on the topic itself narrowing the questions that can be asked, usually occurring for psychology. These observations can also be related to the way the specification describes the content, which can limit the questions that can be asked on a topic, but not all comments refer directly to the specification.

A part of this factor relates to the way that once a topic for the highest tariff questions has been chosen, often based on cycling or time since last appearance, for this topic there are often a limited number of specific questions which can be asked that fit the tariff. So the question almost writes itself.

9. Topic difficulty fits position on paper or question tariff

Specific points about the difficulty of the topic fitting the position on the paper or question tariff are included here. These could be, for example, relatively straightforward questions appearing earlier in the paper or on low tariff questions or particularly stretching topics appearing in a high tariff (synoptic) question.

This factor also includes comments relating to the need to include stretching/basic questions or questions assessing different assessment objectives, and therefore issues of balancing difficulty. This can be considered a sub-category of factor 8 with specific reference to difficulty.

10. Differentiating between candidates of different abilities

Comments related to the question becoming more likely because it effectively differentiates between candidates of different abilities. We have not coded predictions that are just ‘good questions’ as factors, since there is no real rationale, but this factor is a more specific statement, and captures the need to include some questions that promote effective discrimination across the entire ability range and that the selected topic is particularly well suited to this, which may increase confidence in the prediction.

3.2.1.4. Factors related to the logic of whole papers

11. Need to cover part of content not assessed elsewhere/content balance

This factor relates to the selection of questions and topics necessary to construct a coherent, balanced paper, or that a topic has not yet occurred anywhere on the paper but needs to do so. Comments may refer to the need to balance content within a section, or at the level of the whole paper, depending on the paper structure.

Often choice of topic for other questions will narrow down the choice for the current one. There was often a logical process of paper construction observed, where teachers started with one question on a section where it was easiest to narrow down the match of topic to question, and this prediction then restricted the choice of possible questions on one or more other question slots.

This factor also relates to the overall curriculum coverage of the assessment, particularly where the paper covers much of the curriculum - at least the high level topic areas - so most of these main topics have to come up somewhere. Comments against this factor are particularly prominent in the structured history and government and politics papers, where almost all the high level topics on the specification are covered in some form on each paper, and getting the appropriate balance of topics is important.

12. Question effective at assessing a substantial portion of syllabus

These comments relate to good coverage of the specification content by a single question. This may be desirable because it promotes good overall coverage of the content on the paper, or because it provides a stretching/deep question to test the most able candidates. Sometimes this factor will be stated explicitly, sometimes it is implied by the stated links between the content required to answer the question.

13. Need to balance types of questions in unstructured sections of papers

This relates most to the construction of coherent papers in psychology, with a balance of question types, not content coverage. In the predictions this usually referred to the need to include source/data analysis, multiple choice or extended response questions somewhere. Some of these needs are requirements of the specification, other comments reflect having the same kind of balance of question types as the sample assessments in psychology. This factor was not used for history or government and politics.

14. Logical/chronological order of questions

This factor captures the way predictions of questions are influenced by adjacent or nearby questions, where there is some logical flow to the questions. This flow may be related to the way a topic is explored in increasing depth as questions progress, or actual chronological order of topics, particularly for history.

Some papers are structured with a set of questions all falling under one topic area, sometimes with source material. Within the section, question predictions can be constrained by another question prediction in a section – either a high tariff question or the initial source-based question.

This is a much more narrowly focused factor than achieving balanced content coverage, and focusses specifically on the flow of questions with a sub-section of a paper. Explicit comments may be 'this follows on from question a)' or 'because question a) is …..then this will be…' or 'this provides a good contrast to question 3'.

3.2.1.5. Factors related to the age of the syllabus

15. Long-lived spec and likelihood of unusual or random questions.

Some comments suggested that towards the end of a specification lifetime examiners may start to use unconventional questions, perhaps focusing on small sub-parts of a common topic that are rarely assessed, or asking for an atypical type of analysis on a common topic. They may also set questions on topics which may be rather niche parts of the specification content, including areas that are normally considered unlikely to be assessed as students may not be taught them well, or at all, or textbook coverage is limited.

16. New spec with limited past papers means existing questions will not be used

This factor captures comments that questions on past papers or sample assessments are unlikely to come up early in the specification lifetime while there are still topics which have not been assessed yet – so past questions can be ruled out with some confidence to limit predictions. These comments are more specific than question-cycling comments and refer to newness of the specification.

17. Avoidance of new topics in early days of spec

Some comments related to the exam setters avoiding too much new material as they believed this would not be fair to candidates while teachers are still familiarising themselves with the new material. Comments suggested that these new topics are likely to be favoured more when the specification is more established.

18. New topics on spec need to be assessed in the early live papers

The reverse of factor 17 were comments relating to the need to straight away test the new topics on a specification, so examiners can see, for example, that they are being taught well and candidates have absorbed the new material. This requirement needs to be balanced against generally not repeating questions in the sample assessments or practice papers, which often include assessment of new topics.

3.2.1.6. Factors revolving around other resources such as textbooks, sample assessments, availability of sources or topicality

19. Structure of textbook

Textbooks approved for particular specifications were sometimes thought to be influential in that questions would be asked on papers that followed the organisation of the content in the textbook. In a similar way that the definition of the content in the specification can limit questions, textbook organisation could also help question predictions.

20. Topics appearing in textbooks make these topics more likely

The thorough coverage of a particular topic in a popular course textbook means that a question related to this particular topic is more likely as candidates will know it well, will have been well taught, and there is sufficient depth of material available to give an answer.

On occasion this may not be a central topic in the specification, but the textbook coverage makes the topic reasonable to assess. However, comments here are also often tied with "centrality/importance in specification" as they are closely linked – good textbook coverage implies centrality. Material appearing in revision guides also fall under this category, as do Politics Review and Annual Updates for government and politics.

21. Language used in textbooks

Predictions of questions can be influenced by keywords or language used in a course textbook. For example, if a keyword is used in relation to a particular topic in a textbook, then the predicted question can include the keyword related to the particular topic. Also, if the language in a course textbook is favouring one perspective over another, then sources could be selected which align with this perspective.

22. Exclusion of example questions in course textbooks/sample assessment materials

Because the question has been provided as a practice question in a textbook or a sample assessment it is unlikely to ever be used 'as is' on the paper. These comments may refer to instantly excluding these example questions (and sometimes certain content) from consideration, or sometimes to a variant on an example question being required in order to avoid duplication.

23. Use of example questions in course textbooks/sample assessment materials

The inverse of factor 22. Some predictors suggested that because a question has appeared as a practice question in a textbook, students will be familiar with it and it is likely to be used on the paper, possibly in a slightly more narrowly focused form; the practice question will often be quite broad. The same point can occur with questions on sample assessments, in a different form, or potentially they may be re-used when a specification is a little older or coming to the end of its lifetime.

24. Availability of source material for this topic

Certain topics will have a large pool of the appropriate type of sources, be they cartoons, newspaper articles, book extracts, and so they are often chosen – either they make an easy choice for examiners, the materials are good quality or stimulating, or they are a natural choice for the question type. Alternatively, the topic lends itself to the creation of novel or engaging scenarios/sources.

25. Topicality of question, given timelines for paper production

Questions are often influenced by events or controversies in the world and these can be used to suggest questions or sources that may arise. This is particularly true of government and politics, where topicality is a key part of the specifications, but it may also influence psychology questions, particularly those associated with scenarios or sources.

Sometimes our teachers may have underestimated how far ahead of time the questions are written, meaning that some suggestions under this category may have been too recent to appear on the summer 2017 papers. However, these predictions were still coded under this factor.

3.2.1.7. No factors stated

26. Guesswork or no other reason given

This category was used for several types of rationales where no positive reason for a specific prediction was given. Occasionally for a prediction it would be clearly stated that there was no real reason for the choice, or it was just a guess. Alternatively, only the exclusions of other topics might be listed with no justification for the final choice.

Sometimes questions were suggested just with the implicit or explicit rationale that they were good or reasonable questions. Therefore, not every example coded under ‘guesswork’ was actually an uninformed guess, but there were no comments that clearly detailed a factor that was used to narrow down the choice of possible questions or make the prediction.

We acknowledge that some of the predictions that were coded against a factor may also have been guesswork but with a post-hoc rationalisation given. However, the choice of rationalisation chosen can still give useful information.

3.2.2. Analysis of factor coding at specification level

In this section we report the relative frequencies of the factors within each specification. Where there is more than one paper included for a specification, the data is combined across the papers. There is a choice to be made in how we combine the coding of both raters to count factor frequencies. We can either only count instances when both raters coded the predictions against the factor, or alternatively when only one rater coded the prediction against the factor.

The former (both raters) gives a more conservative but robust measure. However, it may miss out on some genuine influences on the predictions. The latter (either rater) includes some less clear-cut coding, where the coders disagreed, but it gives slightly richer data, and includes more coding against less frequently used categories.

Similarly, for the coding of psychology, we could take the lower of the two counts within each section of the paper assuming that both coders agree on those instances, or we could use the higher of the two counts, which we assume to also represent cases where only one person coded a factor.

There are not particularly large differences between the frequencies of the most common factors under the two coding rules, and so we concentrate the analysis on the either rater coding which is shown for each specification grouped by subject in Figure 3. Appendix E includes a comparison of the outcomes from the two coding approaches.

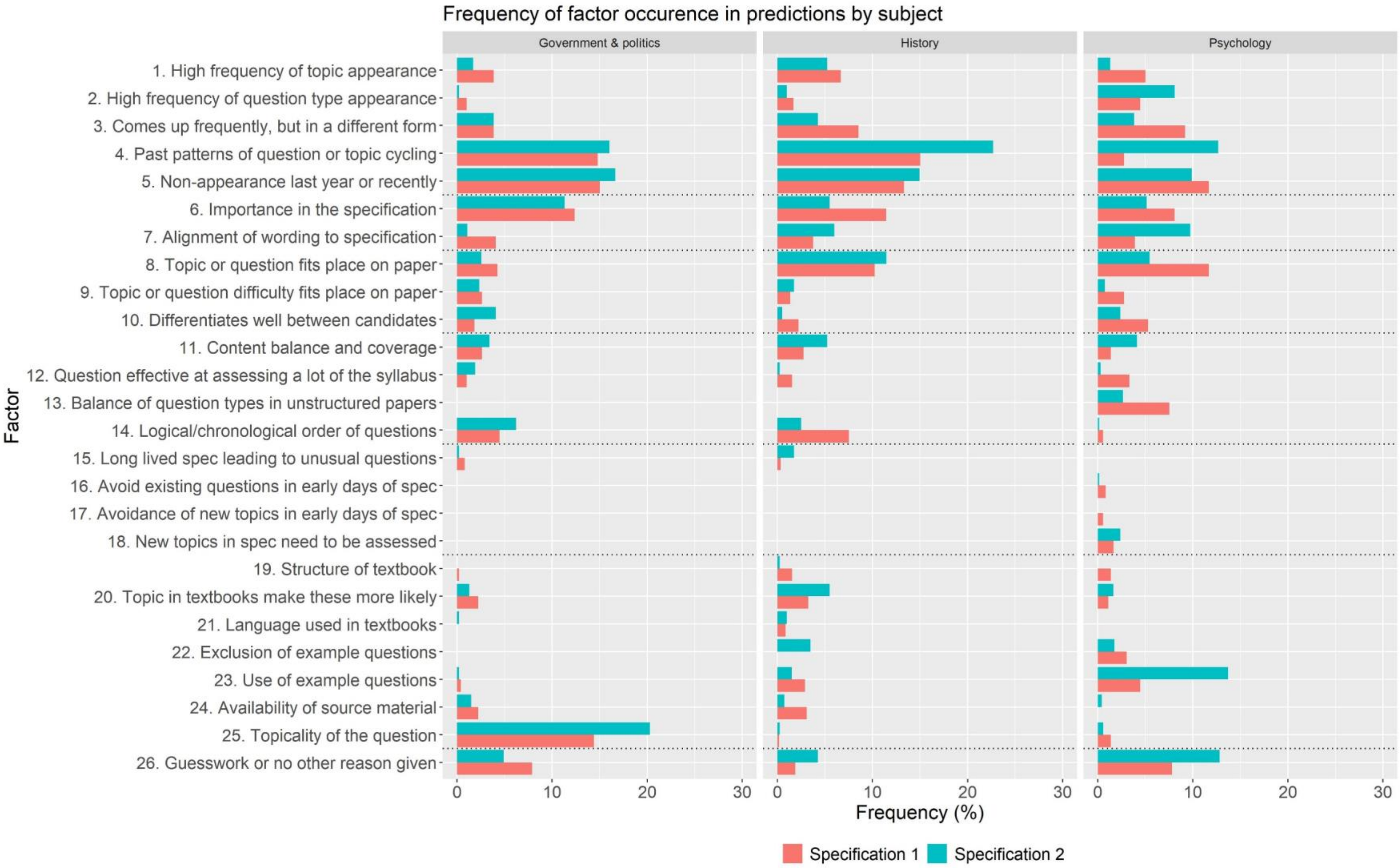


Figure 3: *Frequency of occurrence of each factor in individual predictions made by teachers, broken down by subject in each panel and specification within each panel. The data includes cases where only one rater coded the factor.*

3.2.2.1. Government and politics

There was generally high consistency between the factors identified by teachers across the two specifications. The most frequently identified factors tend to be focused on the importance of the topic in the specification and patterns of question appearance, or non-appearance, on the past papers. Frequency of occurrence of topics was not frequently cited, probably because this was a given as they usually came up.

Topicality was also a very important factor in teacher predictions, especially for specification 2 where it was the most frequent factor with 20% of all coding against this factor. This is consistent with an intention of these specifications to allow reflection on the evolving political landscape.

A variety of less frequently identified factors occurred, often centred on the appropriateness of the content for the specific question, related to the structured nature of the sections within each paper. This structure was reflected in the logical flow of questions factor, which was frequently chosen, since within each section the part b) questions often followed a similar theme as the part a) questions.

3.2.2.2. History

In history, analysis of the past patterns of topic cycling was very frequently used to make predictions, especially for the specification 2 papers where over 20% of coded comments related to this. It was also the most frequent factor coded for specification 1, although at a lower frequency.

As well as frequent reference to topics coming up often, or questions coming up in slightly different forms or not for some time, there were also references to the topic being appropriate to the tariff or type of question (explain, describe etc) asked in the question slot.

Topicality was not really an issue in history, but the coverage of topics in the course textbooks and some reference to example questions were considerations for some respondents. Similar to the government and politics papers, some questions followed on from one another and so the logical flow was used to constrain predictions, more so for specification 2.

3.2.2.3. Psychology

For psychology, a slightly broader range of factors, many relating to what had appeared on the existing sample assessments and practice papers were used, which included use of previously seen questions, or variants of these. Factors given included details on appearance of the topics - in this instance often research studies listed in the specifications - but also the types of questions appearing. This was because for the psychology papers there was a need to pair the type of question to the content area, and a lot of the unpredictability came from the rotation of these two aspects of the questions.

The precise wording used to specify the content was also referred to more than in the other subjects, as the teachers felt the question wording had to follow the specification quite closely.

Although getting the right balance of question types across the paper was sometimes mentioned, this was not that common, possibly due to the way we asked for sets of questions grouped by content area, rather than asking teachers to specify a single final paper where getting the balance of question types right would be more critical.

A number of suggestions were made without specific justifications or just with reference to it being a good question that would test students. The higher proportion of ‘guesses’ probably reflects the uncertainty around a new specification.

3.2.2.4. Comparison across subjects

Topicality was a key difference, arising simply because the government and politics papers are expected to reflect recent events. This doesn’t occur in history and is only a small factor in the psychology predictions, usually occurring in choice of source materials where an issue has been of interest to the media. Predictions for all papers relied heavily on the content in the specification. However, government and politics was influenced more by the importance of the topic.

On the other hand, for history and psychology the wording of the specification was particularly important in limiting the kinds of questions that could be asked about certain topics, and also defining the precise question wording. This is likely to be an underestimate of the actual importance of the specification wording in defining question wording as we explicitly stated that the precise wording of the predictions was not vitally important.

Questions appearing on past papers were important influences across all subjects, with psychology predictions particularly using frequently occurring questions/topics, while for history, particularly the specification 2 predictions, there was a heavy reliance on patterns of question/topic occurrence over time. Example questions were used more to make psychology predictions, unsurprisingly, since no live past papers were available and the sample assessments and textbooks carried more weight. The history and psychology predictions frequently referred to the appropriateness of the topic or question for the place on the paper.

In psychology this referred to the type of response required or question tariff, while for history this was again the type of response required, which was defined by the question number on the highly structured papers. The government and politics predictions referred to this far less frequently because the content areas were all assessed with the same basic question structures.

The importance of the logical/chronological order of questions was important for all of the papers where questions tended to follow a content area within a section of the paper.

For psychology, with largely discrete unlinked questions this was not significantly referred to, although several sections, particularly in the specification 2 papers, had a highly fixed order of question types, this was perhaps too obvious to be mentioned in the predictions. In contrast, creating a balance of questions was more important in psychology predictions due to the generally less constrained nature of the papers.

If we collapse the predictions into their high-level themes, and then combine the two specifications within each subject, the relative frequencies are shown in Table 5. The table confirms the general similarity of factors identified across the 3 subjects, but with some differences.

History predictions were more focused upon the occurrence of questions on past papers, while government and politics predictions more frequently used the other resources category, driven by the issue of topicality and coverage in political review resources. Psychology had a higher level of guesswork or no stated reason, due to the newness of the specification and lack of past live papers.

Table 5: *Proportion of coded comments falling into each high level theme, for each subject, combined across specifications*

Factor category	Government and politics	History	Psychology
Questions on past paper	38.5%	46.7%	34.5%
Specification	14.4%	13.3%	13.4%
Topic/question alignment	8.8%	13.8%	14.2%
Whole paper logic	9.8%	9.9%	10.0%
Syllabus age	0.5%	1.0%	2.8%
Other resources	21.5%	12.2%	14.8%
Guesswork/not coded	6.4%	3.1%	10.3%

3.2.3. Exclusion of topics/questions

Often during the justification for predicted questions, teachers went through a process of elimination of topics. We did not code this as these were not positive reasons for why the selected question would be expected to come up. Generally, the exclusions were very much based upon the pattern of recent occurrences, whether the question/topic or something similar had come up in the most recent paper, or frequently over a period.

We also asked the teachers to fill out a section on the response document detailing what they did not expect to come up, and why. Again, these exclusions were very much focused on the patterns of recent appearances for that topic.

One noteworthy difference was that in government and politics, the exclusions were often based around the lack of current relevance for certain kinds of questions, a frequent example being questions asking about coalition government, which had been common up until that year but had become less relevant following the election of 2015.

3.2.4. Overall observations

We also asked teachers to give any summative comments that they felt were appropriate. These comments fell into several categories, in no particular order:

- observations on success or failure of their own past predictions
- a critique of the rationale for the study on the basis that questions were not predictable, assuming that papers had been chosen for this study because we thought they were too predictable

- pointing out that only going back 4 examination series was not sufficient to confidently make predictions and that a larger set of materials would have been helpful
- listing the major or overarching principles on which they had made their predictions, which largely reflected their most frequently coded factors on the individual questions
- detailing the differences between sections or papers that they'd predicted, including which ones were easier to predict and why
- detailing which type of questions were more predictable and touching on the logical order in which they had generated their predictions – effectively describing the paper writing process that was followed in the later meetings (see Section 3.3)

Finally, all the teachers in this phase also completed a survey on their views relating to predictability. We were interested to know whether the predictability of specifications was a major concern for teachers, and also whether they tried to predict topics and questions that might appear on papers in the future, and how they used these predictions. These responses are analysed and described in Appendix F.

3.3. Phase 2 - Meetings to decide final best predicted paper(s)

Whole question papers were constructed in these meetings, with no alternative questions. The precise wording for each question was agreed, although because the evaluation of the questions was to be based on the underlying skills and knowledge tested by the question, this wording was not absolutely critical. However, the teachers were all keen to make sure the wording was clear and fitted the style of the paper.

Where sources occurred on the paper we only asked for a general outline of the kind of source material and the content area that it would address, rather than specifying it in detail, although sometimes the predictions were quite precise. In all the meetings, a full set of questions and sources was listed for every paper, with no omissions.

3.3.1. Overview of different approaches to constructing papers between the subjects

Predicting questions is a case of trying to second guess exam setters by partially reproducing the process they would go through in writing the next paper in a series. The aim of the exam setters is to construct an entire, coherent paper with the appropriate differentiation of candidates, content and skills coverage, and degree of unpredictability in questions.

Teachers making predictions are likely to follow a similar process, but with less information to hand. Predictability itself arises from how strongly the options for setting questions are limited by the kind of factors discussed in this research, and how much those making the predictions are aware of, and can apply, these factors.

We observed a general approach to narrowing down choices in predictions:

- exclude last year's questions
- see if there are patterns in previous years, while recognising that a question setter will try to introduce a little unpredictability here, perhaps repeating topics or breaking what appears to be a regular question/topic cycling pattern
- see what hasn't come up before or for a while
- start to fill some question slots, probably the higher tariff ones, or ones that are slightly more constrained in choice of topics
- fill all the slots on the paper while ensuring that key topics in the specification are sufficiently tested and overall there is balanced content coverage; skills normally fall out through the constrained nature of the question stems/types ensuring correct skills coverage
- specific questions and their wording will be determined by many factors, although alignment with the specification wording is important

For less structured papers the penultimate step is not about filling all question slots, but looking at past patterns to generate suitable questions that in total add up to the right number of marks with suitable content and skills coverage.

It is worth noting that the information we received on the prediction templates did not always describe the whole paper production process since this was not the intention of the documents. Certain aspects were taken as assumed and most participants did not describe them explicitly, such as the exclusion of any questions that came up in the previous year. Although in some cases participants noted that some of these questions may come up in a different form, it was implicit that examiners would not re-use 2016 questions in exactly the same form.

This then provides the first step in the process – exclusion of last year's questions. This was reflected in the responses we received in the 'what will not come up and why' section.

However, from that point on, the process did seem to vary a little between papers, largely due to 2 factors – the age of the specification and the structure of the papers in terms of a fixed or variable set of question types or tariffs. Due to similarity between all the papers within a subject, the process is best described at the level of the 3 subjects.

3.3.1.1. Government and Politics

The content for each government and politics paper is divided into 4 main sections, each of which has its own dedicated section on the paper, mostly having 3 questions. Within each high level section the content is divided up into usually 2 sub-areas. Within each topic area, there are sometimes a limited number of extended questions that can be asked.

This means that the 25 and 40 mark questions are somewhat constrained, initially by the likelihood of some kind of alternation of the two main topic areas, and then when one of these is selected, by the questions that can be asked within it.

Unpredictability comes from the precise questions asked, where changing the sense of the question - advantages or disadvantages, for example - or the scenario it applies to - which political party, for example - is used. Sometimes cross-links between the materials on the 4 sections of the paper (high level content areas) can also be used for the higher tariff questions, to add variability.

A lot of question selection is therefore based on an analysis the previous patterns of the areas coming up, particularly on the 25-mark questions. Often then everything else within a section of a paper falls out of the 25-mark question choice, with the second main topic area in a section usually being tested in the 2 lower tariff questions in a section.

Generally, the lower tariff questions, being narrower in scope, are less predictable, and this is particularly true of the 5-mark part a) questions that are often about defining a term, where it was usually a case of just picking from a list of key features in the specification that hadn't been tested for several years. In all question choices, topicality was a concern, where political events in the appropriate time frame could drive question choice, both for the high tariff questions and also for the choice of sources.

Overall the structure of the papers and specification content is constrained and predictable, but the precise nature of the questions, such as what they require in the response, is varied to limit the scope for teaching to the test. Added to this is the topicality of the specifications, which means that having thorough up-to-date knowledge will be rewarded, so the specification is not entirely fixed.

3.3.1.2. History

These papers have a fixed structure of tariffs and question types, with the paper divided into a set number of sections. There are generally the same number of high-level content areas in the specifications as there are paper sections. Therefore content areas get cycled round the sections, and so predictions are largely a two part process.

The first stage is to decide which content area will be allocated to which section. This is based largely on an analysis of past patterns of cycling of content areas, which is not always a rigid cycle, but appears to contain some randomness.

After allocating content areas to sections, the topics within an area are spread out around the available set of questions, with specific questions written to fit the tariffs and question types available. Past patterns are considered closely, and suggestions may be influenced most by an occurrence of the content area in that paper section some years earlier.

But probably the largest factor is that some parts of topics lend themselves to particular question types. There are limited questions to ask about some topics. Some lend themselves to extended explain/discuss questions, while others are too small for this and fit lower tariff questions. The specification wording can be fairly closely linked to the type of question that can be asked about that topic. This means that there is not total freedom in assigning topics to questions, and these constraints can be used to narrow question choice.

When constructing a set of predictions, often the starting point is the highest-tariff questions. Assignment of topic to this then facilitates the distribution of topics to different questions, as there are a limited number of key topics within each high-level content area, and many of them tend to be assessed in some form on every paper due to their importance to the subject.

Once one specific question in a section has been decided on, other questions may therefore fall out quite easily, at least at the topic level. Even the use of sources doesn't make topics unpredictable, as some topics lend themselves particularly to sources through the amount and/or quality of sources available, and there is therefore slightly constrained cycling of topics on the source questions.

Overall topics are often quite predictable, but a variety of questions can be asked on each topic to moderate predictability, although sometimes the specification content can limit this.

The amount of content for each section affects the degree of perceived predictability. For example, it was stated several times by specification 2 predictors that the Germany paper was less predictable than the Cold War paper, because there were fewer substantial key topic areas to slot into the Cold War paper, particularly the higher tariff question slots, meaning that there were less permutations across all the question slots.

However, although most or all high-level content areas will appear on each paper, at a more detailed level the amount of content covered in each paper is still much less than the specification content, so that there is still a lot of uncertainty around specific questions that will come up.

3.3.1.3. Psychology

Both the psychology papers have flexible question structures within each section. However, they need to include certain types of questions across the sections: a distribution of research methods and also extended response questions. The sections themselves and the content they assess are fixed.

So the predictions focus a lot on getting the spread of types of questions both within a section and across sections correct. This leads to a lot of matching of topics to question types, as in many cases each part of the content lends itself to particular types of questions, or depth of answers.

Predictions are not really constrained by the structure of the paper, but are limited by matching topic to question type and using the patterns and recent questions in the previous papers to narrow down what may come up. Recent or frequent appearances of a topic-question type pair can be used to narrow down the type of question most likely for a topic.

In psychology it is very likely that certain topic areas will come up in some form due to their importance, but this is not completely guaranteed since for specification 2 there tend to be fewer questions than core topic areas, and within each section of the specification 1 paper not all topic areas can be assessed and there is variety in those that are tested each time.

For some sections on these papers, particularly specification 2 sections B and C, there was actually a very high predictability of the question type, but this was in conjunction with a very low predictability of the content area. Repeated types of questions within a section were applied to different material each time and so overall predictability was not a problem.

In many ways this is similar to what occurs in English papers, where the skills required are well known by candidates, but the material on which they will have to apply them is unknown.

One additional factor that we had not provided the teachers with were the corresponding AS papers, particularly the summer 2016 AS paper. So although questions on the sample assessment materials and practice papers were generally used to narrow down likely questions, exclusion of questions on the AS paper was a powerful predictor, as many of the same candidates would sit the AS and A level papers so most teachers thought questions would not be repeated across these two papers.

Being a reformed qualification with a lack of live past papers there was also a high reliance here on statements in the specification. Questions tend to be prompted by the key points and also skills detailed on the specification. This may change over time as the set of past papers grows.

3.3.2. Analysis of factor coding at specification level

Following the meetings, the recordings were analysed and coded against the same set of factors as the individual teacher prediction documents. In this instance, due to resource constraints, only a single researcher carried out the coding. However, given that this was one of the researchers who coded the teacher prediction documents, the same consistency in applying the coding scheme as described in Section 3.1 would be expected.

The predictions for the exam board taught by the team - the ‘home’ prediction - and the second specification in their subject - the ‘away’ prediction - are shown for each subject in Figure 4 to Figure 6. Often many of the same factors were mentioned in the meetings as had been referred to in the individual teacher predictions.

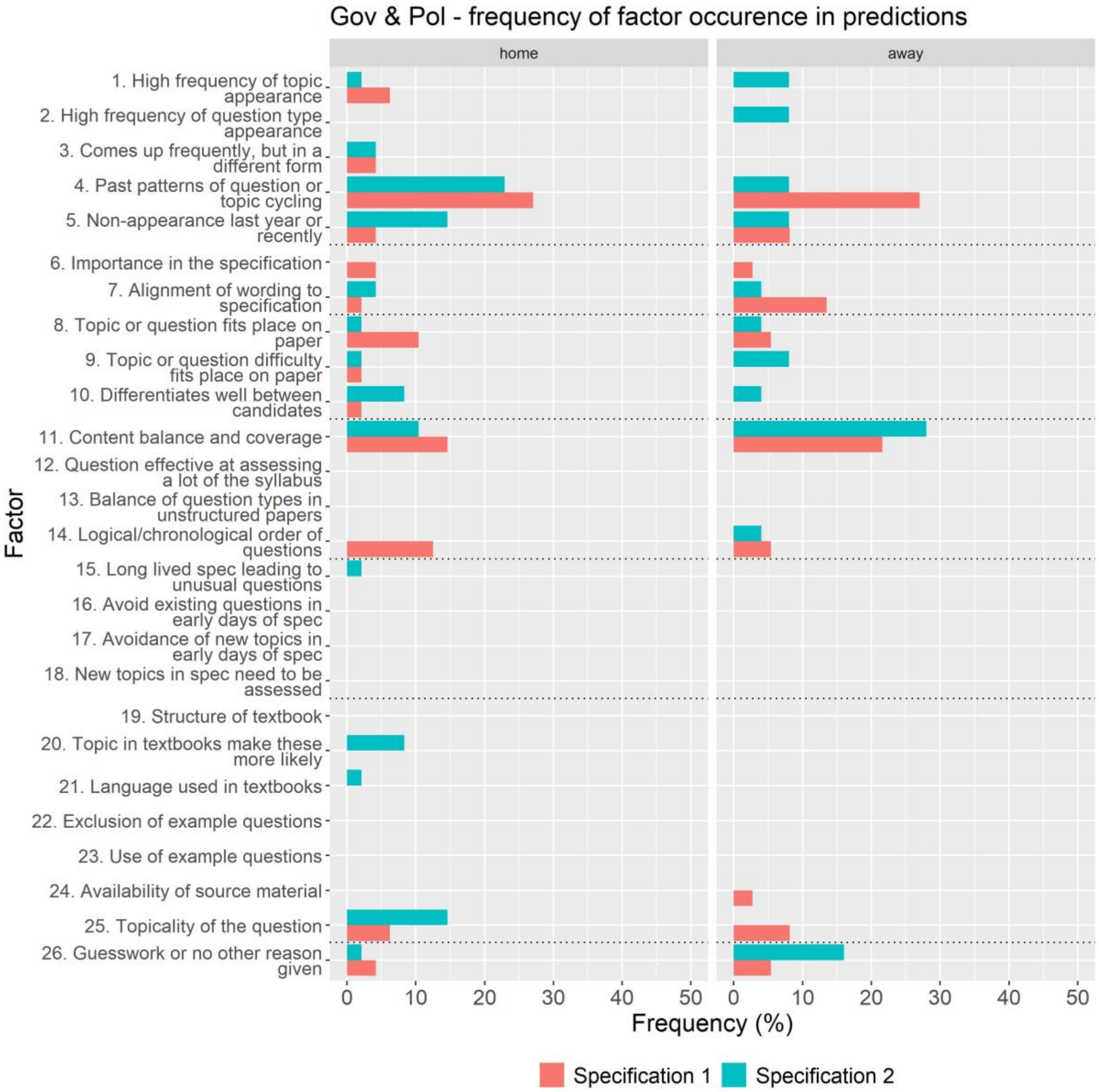


Figure 4: *Frequency of occurrence of each factor in meeting predictions for government and politics. Specifications are distinguished by colour with predictions for 'home' papers in the left panel and 'away' papers in the right panel.*

Figure 4 shows the coded factors from the two government and politics meetings. Compared to the factors identified in the teacher predictions (see Figure 3), past patterns of question cycling was frequently referred to, particularly for the specification 1 papers, both 'home' and 'away'. Similarly, content balance and coverage was frequently referred to, particularly for the 'away' predictions. It may be that for the 'home' predictions familiarity with the papers meant that this arose automatically from their predictions and did not need to be explicitly stated.

Alignment to the wording of content in the specification was also used more for 'away' predictions, particularly the specification 1 'away' paper. With no experience of teaching these specifications, the participants had to rely more on the provided materials. Topicality was less frequently referred to in the meeting compared to the individual teacher predictions, while guesswork was used more, particularly for the ‘away’ predictions – when the ‘home’ paper predictions were coded here this often was an implicit understanding of the logical flow of questions or content balance which was not stated out loud.

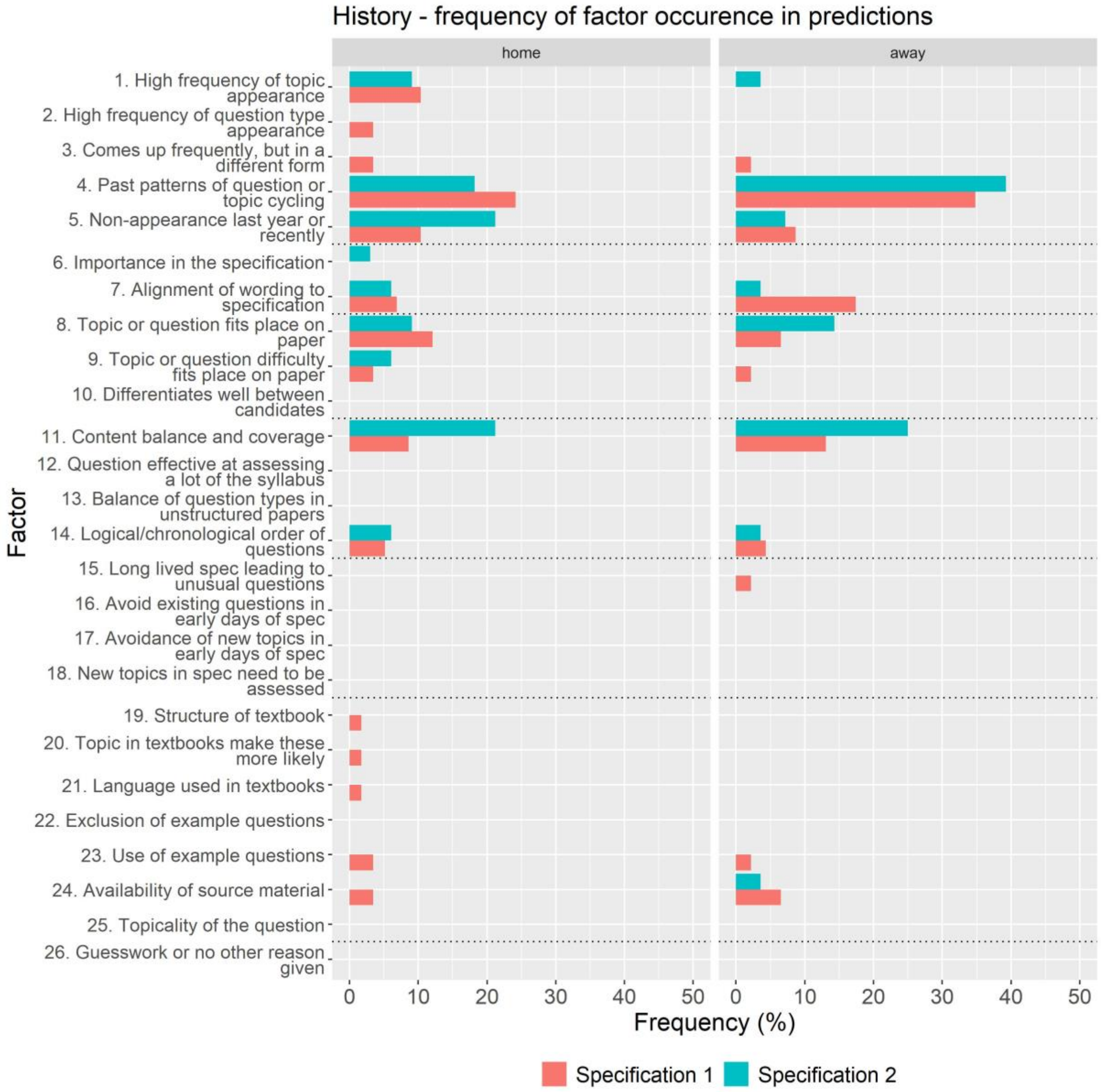


Figure 5: *Frequency of occurrence of each factor in meeting predictions for history. Exam boards are distinguished by colour with predictions for ‘home’ papers in the left panel and ‘away’ papers in the right panel.*

The pattern in the history meetings (see Figure 5) was similar to that seen for government and politics, with past patterns of content cycling, content balance and coverage and to some extent the alignment of question wording to the specification content all being frequently mentioned. These were all similarly frequent in the individual teacher predictions. Content balance and coverage was mentioned more frequently for the specification 2 papers, by both groups of teachers when they were both ‘home’ and ‘away’ papers. Non-appearance the previous year was more influential for the ‘home’ paper predictions than the ‘away’ ones.

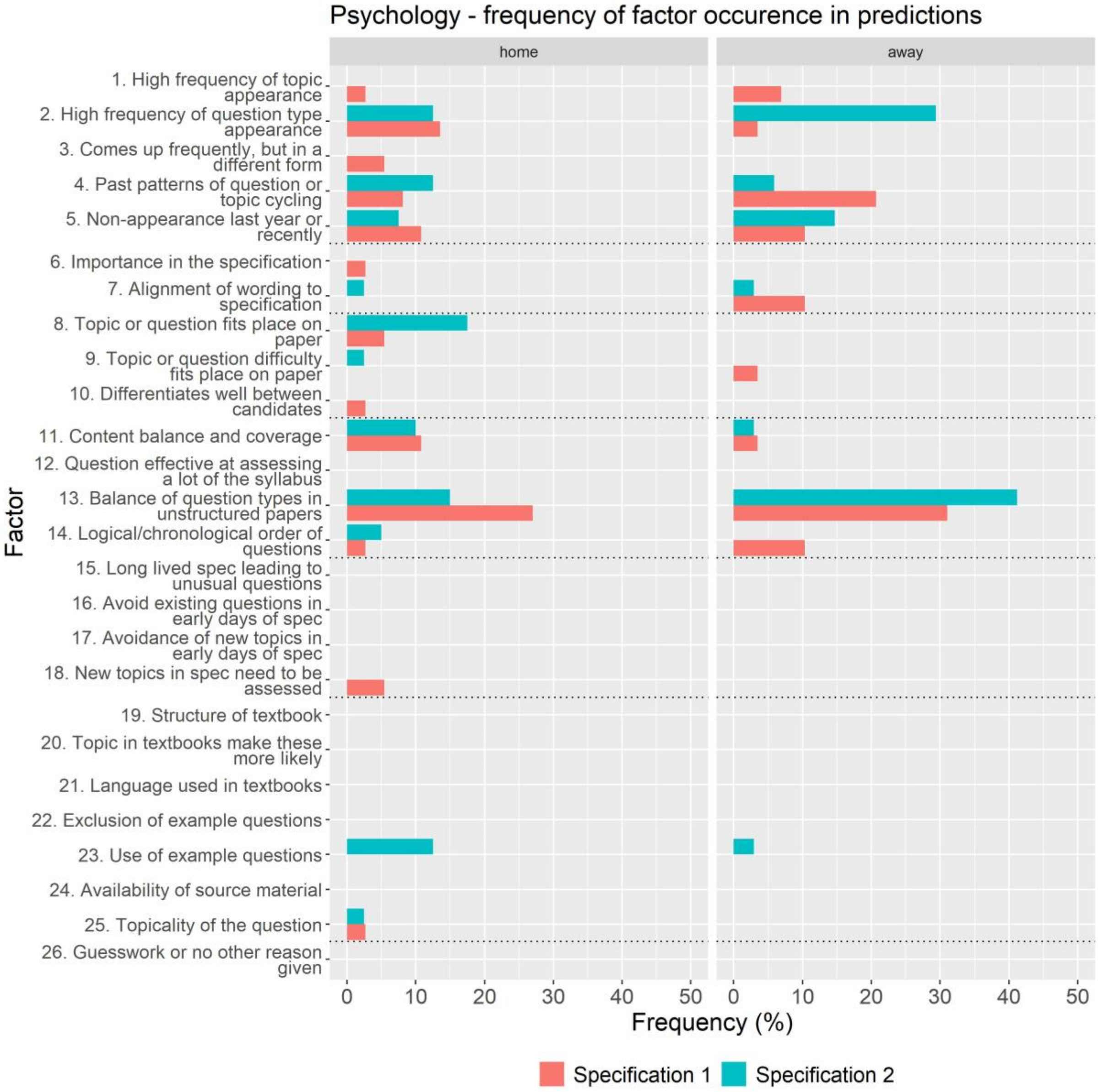


Figure 6: *Frequency of occurrence of each factor in meeting predictions for psychology. Exam boards are distinguished by colour with predictions for ‘home’ papers in the left panel and ‘away’ papers in the right panel.*

In the psychology meetings (see Figure 6), there was a focus upon the balance of questions types across the paper for both specifications and both ‘home’ and ‘away’ predictions. This is to be expected given that the paper sections are quite unstructured. This was even more frequently mentioned when the ‘away’ predictions were being made. Although ‘home’ board predictions sometimes used the appropriateness of the content for the question type, this was not mentioned when making ‘away’ predictions, perhaps because the predictions from the first phase were being used and so this aspect was assumed when picking from these previous predictions. Past patterns of occurrences were studied in all cases, notably the frequency with which question types appeared for the specification 2 away predictions.

Across all 3 subjects, in contrast with the individual teacher predictions in Figure 3 there was a lot less reference to frequency of topics coming up in the meetings, and factors around textbooks and example/sample paper questions. This probably reflects the slightly different focus during the meetings, as individual teacher predictions were reviewed and selected. Discussion was focused on attempting to select questions from predictions to produce coherent sections on the papers rather than discussing the rationale behind all of the predictions.

3.4. Phase 3 - Expert ratings of overlap of predicted and live question papers

For each question on the live papers, the experts gave a rating out of 10 for outcome space overlap. In one instance an expert gave two separate ratings for content overlap and skills overlap, which in consultation with the expert were weighted 3 (skills): 7 (content) which meant that a predicted question which required the same skills, perhaps using the same kind of command words, but covering entirely different content would be rated 3 out of 10 on the overlap measure. We will return to the question of skills and content overlap in the final discussion.

To generate a metric of how well the live summer 2017 paper had been predicted, we rescaled the ratings into the range 0 to 1 to turn them into a proportion overlap, and then weighted all the different question overlaps by multiplying them by their tariff. These summed weighted proportions were then divided by the total marks on the paper and converted into a percentage by multiplying by 100. Equation 1 shows this calculation,

$$O = \frac{\sum_{i=1}^{n} o_i m_i}{\sum_{i=1}^{n} m_i} \times 100 \qquad (1)$$

where O is the total paper overlap, n is the number of questions on the live paper, o_i is the overlap rating for question i in a range (0,1), and m_i is the tariff for question i.

This gives an overlap score for the paper as a percentage of the outcome space of the whole live paper that was correctly predicted. Higher overlap indicates that a higher percentage of the knowledge and skills required to answer the questions on the live paper were contained within the outcome spaces of the questions on the predicted paper, meaning that the predictions were more accurate.

However, note that this figure does not in any way represent the percentage of whole questions that were perfectly predicted. In only a few cases were questions perfectly predicted with an overlap rating of 10. Far more frequently a lot of questions on the live paper had a moderate degree of knowledge and skills overlap with predicted questions.

Figure 7 shows the calculated overlap for each specification. Where more than one paper from the specification was included in the study they have been combined, with all of the questions entered into equation 1 and divided by the sum of the marks on the two papers, so that if one paper had a higher mark total it would contribute more weight to the total. To start with we have combined the two sets of predictions from the 'home' and 'away' teams, so we include both experienced, and naïve prediction groups for each specification. We will consider the accuracy of the different groups below.

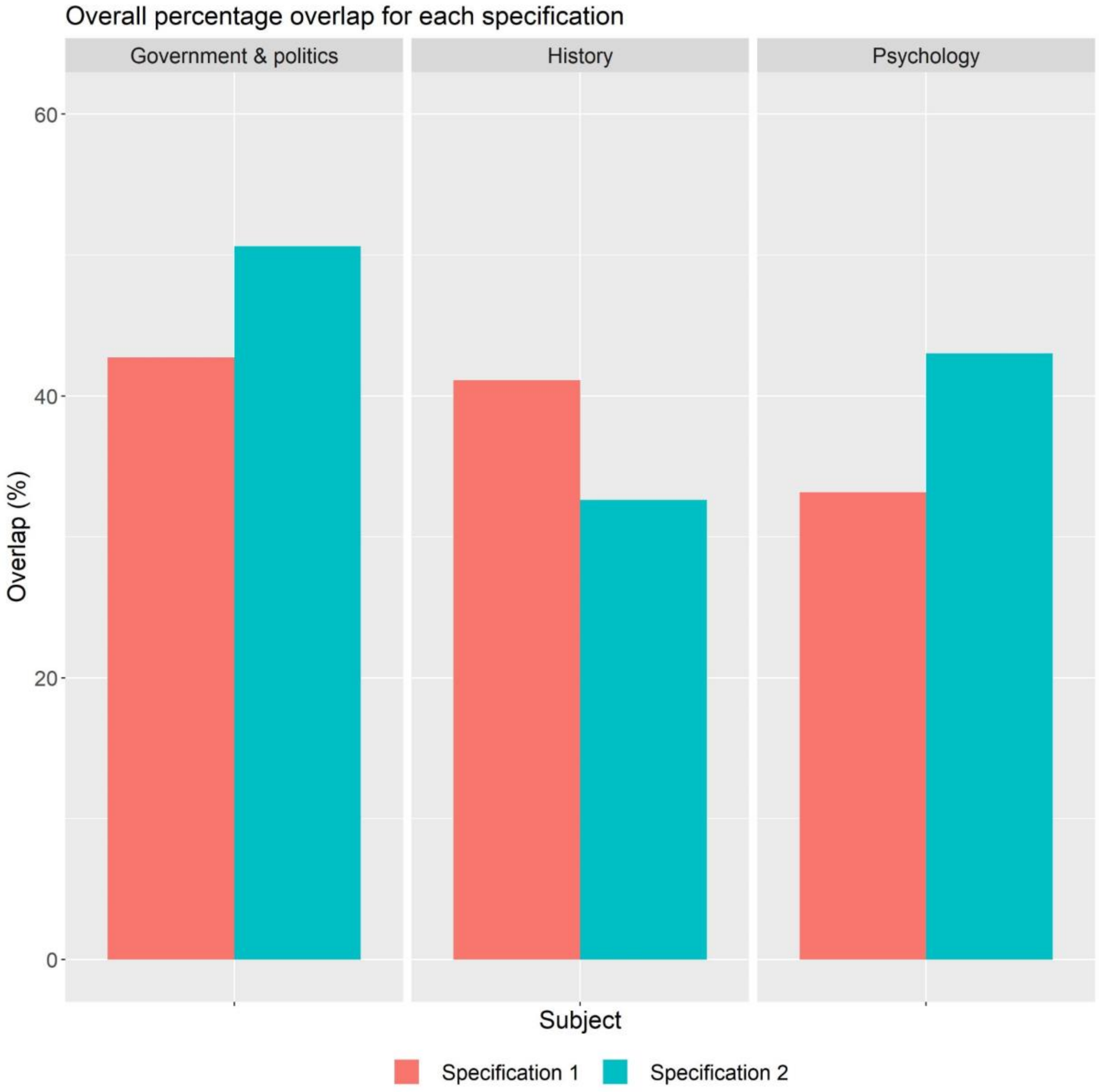


Figure 7: *Calculated overlap percentage for each specification, averaged across all papers and both the 'home' and 'away' team predictions. Specifications are distinguished by colour.*

In each subject, one specification was more accurately predicted than the other, by around 8-10%. Given that this data was based upon just two sets of predictions for each specification then comparisons within subjects are indicative only. Comparisons across subjects showed that the government and politics specifications were predicted the most accurately, with the two specifications the 1st and 3rd most accurately predicted. If we average across the two specifications, government and politics had 46.7% overlap, psychology had 38.1% overlap and history had 36.9% overlap.

The question of whether thorough knowledge of teaching and preparing candidates for assessment in a particular specification affects the accuracy of predictions is address in Figure 8, where we separate the overlap percentage for the ‘home’ and ‘away’ team predictions. For history, averaged across specifications the ‘home’ predictions (38.7%) were more accurate than the ‘away’ ones (35.0%), while for both government and politics and psychology, the ‘away’ predictions were more accurate (government and politics: ‘home’ 42.3% vs ‘away’ 51.0%; psychology: ‘home’ 32.6% vs ‘away’ 43.6%). Averaged across all 3 subjects the ‘home’ predictions were less accurate (37.9%) than the ‘away’ predictions (43.2%).

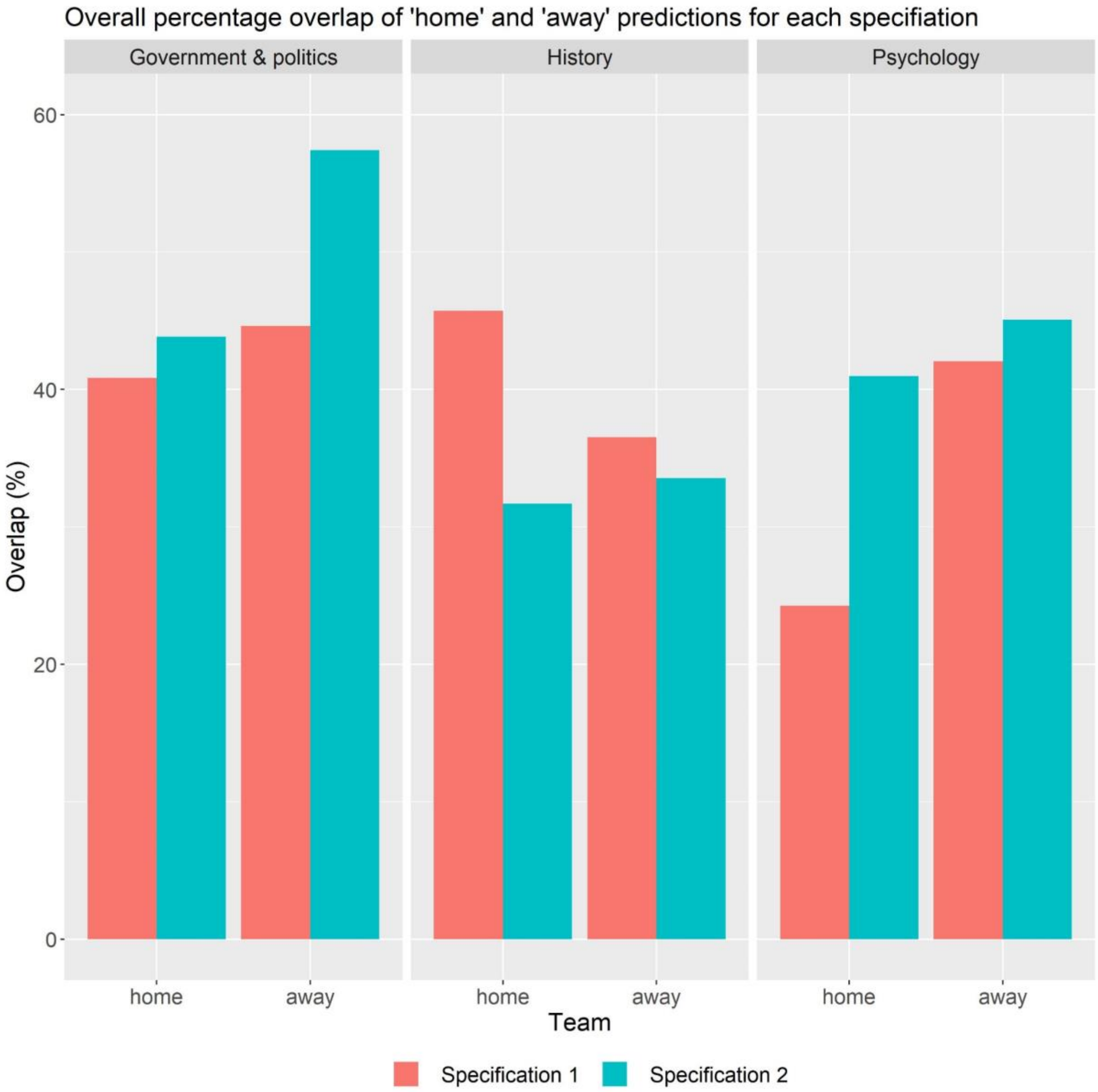


Figure 8: *Calculated overlap percentage for each specification, averaged across all papers, showing both the ‘home’ and ‘away’ team predictions. Specifications are distinguished by colour.*

Given that this was a balanced within-subject design, with the same groups making both 'home' and 'away' predictions this advantage for 'away' predictions is a noteworthy finding. The same materials were available in both exercises, including the full set of individual teacher predictions.

We can see two possible explanations here. One is that the teachers that attended the meetings were below-average in their independent predictions, but favoured and argued for their own possibly sub-optimal predictions, while in the 'away' condition they were able to objectively pick the best suggestions from those provided. This does not seem likely to have arisen through the random recruitment of the 23 teachers, out of the original 55, that attended the meetings.

The alternative is that a thorough understanding of the specification, meaning a long awareness of the make-up of papers, has not helped make predictions. Effectively too much information has hindered them in making predictions. They may, for example, have come to this study with preconceptions of their own specification, for example thinking that there are patterns in the papers that are not actually there. When making 'away' predictions purely on the basis of the provided predictions, specification content and recent past papers, they would naturally be free of any preconceptions or false assumptions.

Finally, we can split some of the whole specification data in Figure 8 into the individual paper overlap, as shown in Figure 9. Less confidence should be placed on differences between papers within a specification, given that these percentages are based on a smaller set of questions and fewer predictions; the papers are smaller where there is more than one selected for a specification. However, specification 2 in government and politics shows generally high prediction accuracy across both papers, while for specification 2 government and politics paper 2 appears to be harder to predict, especially for the 'home' predictions. In history, specification 2 shows a large and consistent difference across the papers.

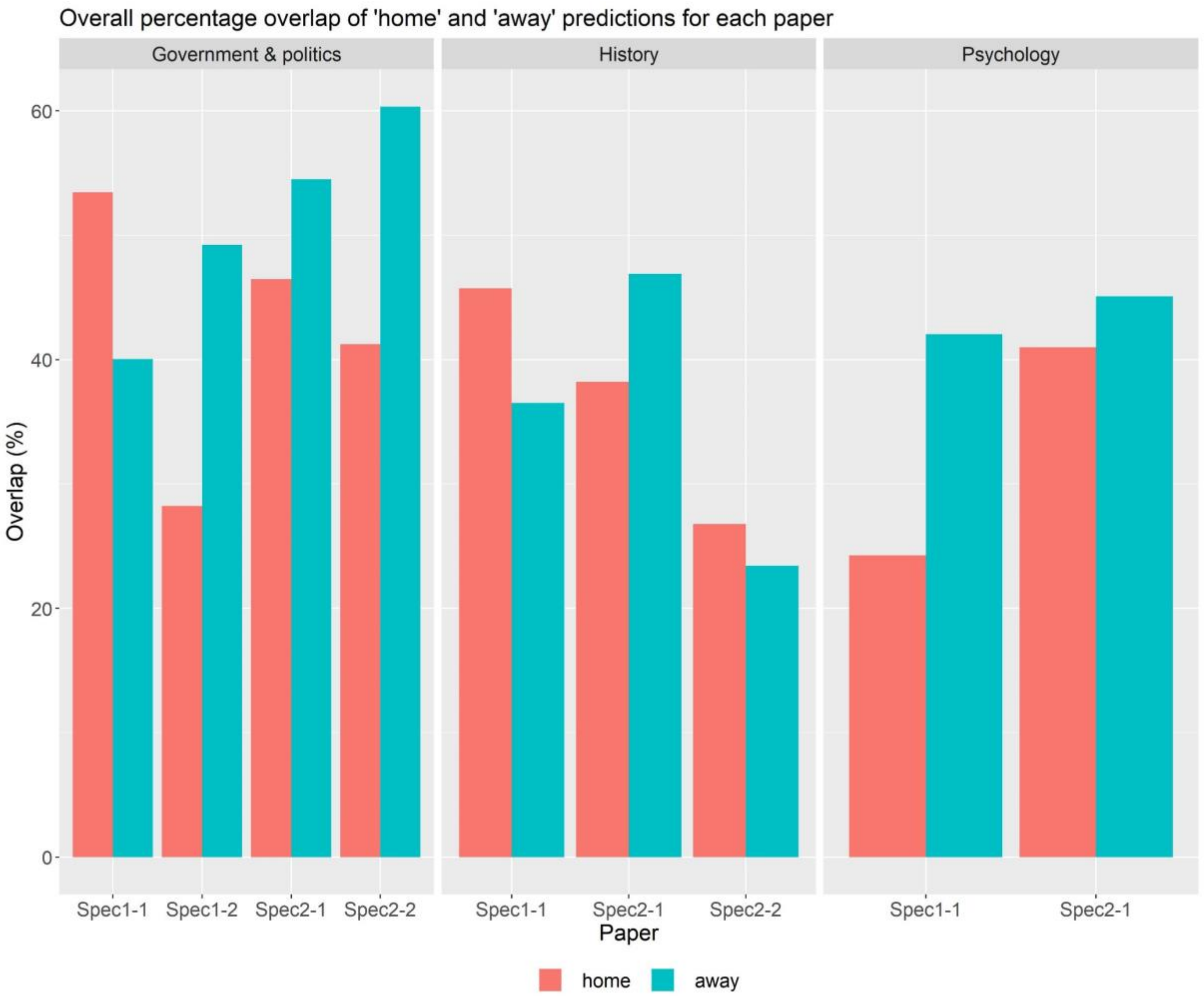


Figure 9: *Calculated overlap percentage for each paper in the study showing both the ‘home’ and ‘away’ team predictions.*

Overall, it is perhaps surprising that the psychology papers did not come out as less accurately predicted than the papers for the other two subjects, given that these were the first live papers in the series, and they had a less constrained structure than those in the other subjects. There are two factors to consider here.

One is that although no previous live A level papers had been sat, the first AS papers in psychology had already been sat, and these had substantial similarity in content to the A level papers, and could thus be used as strong predictors to narrow down question choices.

The second factor is that there was more of a tendency for questions assessing similar skills but in very different content areas to be rated as overlapping by the psychology subject experts than the other subject experts. Perhaps because the history and government and politics papers were completely constrained in the question tariffs and the types of questions asked at each slot on the paper, the history and government and politics experts were much more focused on content areas.

The conclusion here is that the actual content that may come up and the precise question that may be asked on it, is highly unpredictable in the psychology papers, but the types of questions that are likely to come up, and the corresponding skills they assess, are actually quite predictable. The way the outcome space overlap has been operationalised leads to psychology predictions being rated as having similar predictability, or overlap, to history. It is clear though that the government and politics papers were in general the easiest to predict.

3.5. Phase 2 and 3 interaction - Factors identified in the meetings for predicted questions that came up in some form on the summer 2017 papers

The final stage of analysis was to take the expert ratings of live and predicted question overlap and identify which of the predicted questions were good predictions. The factors mentioned in the meetings as influencing the decision to pick questions can then be partitioned between accurately predicted questions and inaccurate predictions. Note that this analysis is based upon quite a small number of predicted questions, and so the results are only suggestive of which factors led to better predictions.

The task of the experts had been to identify live questions for which a predicted question was a close match, so we had to work backwards to identify the correctly predicted questions. The criteria for identifying correctly matched questions were:

- identify questions on the live papers for which the overall overlap, across the 2 or 3 experts, was 50% or greater
- identify the predicted questions that were considered the best match to the live question
- if all experts selected the same predicted question as a match this was a correctly predicted question
- where the experts matched the live questions to different predicted questions, we only counted a predicted question as a match where the expert(s) picking it gave an individual similarity rating greater than 50%

Over all 3 subjects, about 25% of the predicted questions were categorised as accurate predictions. The data were combined across the two specifications in each subject due to the relatively small dataset, so we present the factor frequency counts for the correct and incorrectly question predictions at subject level in Figure 10.

Again, it is worth repeating that this analysis is based upon quite limited data, as for each subject only 18 to 25 questions were categorised as correct predictions. The figure shows frequency counts for factors coded against the correct predictions in blue and the incorrect predictions in red - which are treated as negative counts to visually separate the prediction types.

Clearly there are more factors coded for incorrect predictions, but the measures of interest are both the absolute frequency of factors occurring in correct predictions and the relative frequencies of correct and incorrect predictions for each factor.

There are almost no factors which led to more correct predictions than incorrect ones, particularly where there were more than one or two occurrences of that factor. For history, high frequency of topic appearance, and a question coming up frequently but in a different form, often led to accurate predictions.

For psychology, alignment to the specification wording was referred to 3 times in accurate predictions. Other cases were generally one-off instances of the factor, although it is interesting that for history, all 3 predictions based upon factors associated with textbooks were accurate.

Most other factors that did generate accurate predictions also produced a larger number of inaccurate predictions, so their underlying predictive value may be debatable – a large number of predictions associated with a factor are always likely to produce some accurate ones by chance. However, past patterns of topic cycling occurred for all 3 subjects, although in all cases it also generated a lot of inaccurate predictions, especially for history.

For psychology, getting the balance of question types across unstructured sections produced some accurate predictions, as did the high frequency of appearance of particular question types. For history and government and politics, consideration of content balance generated some accurate predictions amongst a larger number of inaccurate ones.

Overall, the analysis here did not suggest any factors in particular were strongly associated with correct predictions, although the limited size of the dataset must be acknowledged. This finding perhaps underlines how making predictions judged by experts to be correct predictions is quite a difficult task.

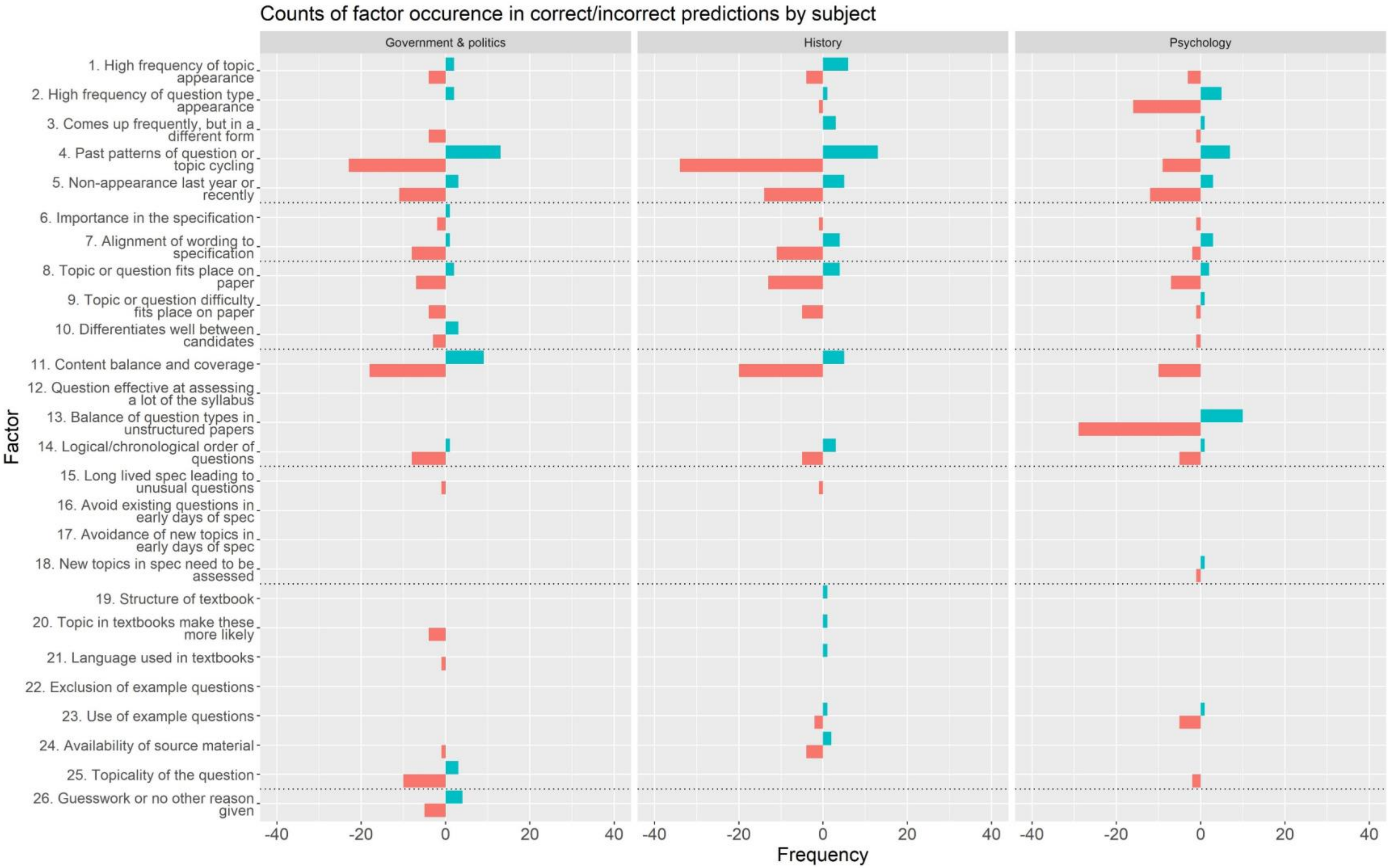


Figure 10: *Frequency counts of the occurrence of each factor in correct and incorrect question predictions made during the meetings, broken down by subject in each panel. Positive counts in blue indicate correct predictions, and negative counts in red indicate incorrect predictions.*

4. Discussion

Before considering the findings further, we must acknowledge that this was a limited sample of subjects and papers, with mainly constructed-response questions ranging from short answers to extended essays. If the exercise was carried out in other subjects, additional factors might be identified, and the relative frequency of the factors identified here would almost certainly differ.

For example, in mathematics and the sciences, papers consist of larger numbers of questions that sample most of the content, which could lead to predictability, but this awareness offers little or no advantage - knowing that simultaneous equations will be assessed offers no benefit if you don’t know how to solve them.

At the other extreme, English literature papers offer substantial optionality, but with only perhaps 2 questions being answered by each candidate. Candidates will know that there will be questions on the paper for the literature they have studied, but there should be scope for varied questions on each work of literature to control predictability.

For the 3 subjects studied here, we saw differences in the approaches taken to construct the predicted papers, and the predictions also varied in accuracy between subjects, but there were similarities in the factors identified as influencing predictions. Therefore, although the factors listed in this report and their relative frequencies may not be representative of all subject assessments, they do provide rich data for the exam boards, and specifically test developers and item writers to consider when trying to control predictability.

Given that these factors are likely to lead to quite predictable questions, particularly if they act in combination, writing questions that do not all follow these factors will lead to less predictable tests. Some of the main factors are expanded on in the following section.

4.1. Factors influencing predictions and their consequences

Previous work has highlighted similar high-level issues to some of those identified here. Murphy et al (2012) summarised the key predictability threats identified in the Ofqual (2008) report as follows:

> 1. The structure of the paper. The key issue was whether the examinations allowed considerably less coverage of the prescribed content at little risk to the candidate.
>
> 2. Relying on the wording of the syllabus content, which was itself over-specified. This in itself reduced the flexibility of the question setters. In order to align closely to the specification, questions were seen to use the same wording. This, in some cases, led to formulaic and predictable questions.

> 3. A predictable cycle of choice of topics. This would make it reasonably easy for teachers to 'spot' topics and prepare their students to deal with the questions on them.
>
> Murphy et al (2012, p.2)

We will expand and extend these existing categories in the following sub-sections.

4.1.1. Questions on past papers

The largest influence on all predictions were the occurrence, or otherwise, of questions on past papers. Although truly fixed cycles were rarely, if ever, identified, the process of predicting questions often started by ruling out or marking as unlikely recently asked questions, and looking for patterns in the way topics appeared in different places or as different question types on the papers.

A degree of unpredictability is vital here, and in fact sometimes a somewhat similar question appearing in adjacent years can be a good thing if this is done infrequently and irregularly, as this topic will have been ruled out by teachers and is unlikely to have been prepared. Re-use of questions assessing the same outcome space should always be approached with caution though, even where the question wording is changed.

As a specification ages, teachers may be able to spot repeated questions and prepare students to answer them. An organised approach to cycling topics around the papers is vital, to avoid fixed patterns which are easy to fall into, and some kind of mapping exercise along the exam series should be carried out.

4.1.2. Specification content

A factor contributing to restricted question possibilities in all 3 subjects was the detail of the content in the specification. Phrases from the specification are often replicated in the examinations because this close alignment is perceived to be fair – the students should have been taught this area of content in the form assessed.

This aspect of the specification wording was generally used late in the question predicting process, when narrowing down the precise wording of a question. However, over-precise definition of the content can mean that for particular content, the range of questions that can be asked is limited, if it is believed to be unfair to ask questions that are about the content more broadly than it is stated.

Questions that do not follow the content specification closely, or which fall outside of the terms used in the specification may potentially still be good questions. They may require the candidate to do more than regurgitate memorised facts, to apply their broader knowledge and skills to construct a coherent answer.

Although in our study these kind of questions were occasionally thought to be possible at the end of the lifetime of a specification, these are the questions that are likely to invite controversy and suggestions of unfairness on social media or in the mainstream media. They are likely to be avoided by candidates, which can lead to optional routes not being chosen by candidates to avoid the unexpected question.

We frequently saw and heard potential questions ruled out in the meetings due to their lack of alignment with the specification content, or even being explicitly judged outside the specification content because the specification wording did not include particular descriptors for that topic. Although issues of fairness and clarity require the content to be clearly delineated, care needs to be taken with the wording of content so as not to overly narrow the types of questions that can be asked about each content area. Over-specific wording will potentially limit student learning, as teaching of particular topics can be narrowed by only focusing on the questions or statements given, rather than exploring all aspects of the topic.

Content definition also interacts with past questions on papers. If a topic area has not been assessed for some time and may therefore be due to be assessed, but the content limits the questions that can be asked, then this may lead to a highly predictable question, and even potentially re-use or adaptation of a previous question.

4.1.3. Textbooks

Textbooks, revision guides and other publications associated with the specifications were also frequently used to predict questions. Sometimes these predictions revolved around example questions provided in the approved textbooks, but more frequently the quality of coverage of a topic in course materials was key, in terms of the way the material was structured and the precise detail provided. This was particularly so for the history papers.

If the approved textbook did not provide material sufficient to answer a particular question this was thought to make the question unlikely even where it clearly fell within the content. Ideally any textbook should provide coverage of the entire specification content to a level of detail that is sufficient for any assessment within the parameters of the specification, and this includes less central or core content, and should not just concentrate on the popular or core content.

Fundamentally, all content in the specification should be assessable at the appropriate level of depth, and textbooks and course materials should not act as additional restrictions on the questions that can be fairly examined.

4.1.4. Sample questions

There was a belief among our participants that questions on sample assessments would not appear in live exams, at least in the first few years of the specification, and even later in the life of the specification they would be unlikely to appear in exactly the same form. Sometimes this could potentially lead to an area of content never being assessed, particularly if a sample question was practically the only question that could be asked around that content due to the wording of the content in the specification. This might lead to tension whereby there is a growing need to assess the content during the life of the specification, but there is a desire not to repeat a question from a sample assessment, and at what point does that re-use become acceptable?

Careful choice of questions on sample assessments is important, to avoid closing off certain parts of the content from being assessed, particularly in the first few years of the specification. Of course, this must be balanced against the need for the sample assessments to be truly representative of the live papers. Less constrained wording in the content specification may help mitigate this problem, allowing content assessed on sample papers to be included again, perhaps in a slightly different form.

The reverse of this issue around content is that the form of the questions on the sample assessments can be highly predictive of the type or wording of questions that will appear in the live series. The questions on the sample assessments may actually limit the scope for question writers to explore alternative forms of questions, even where these may be valid ways to assess the assessment objectives. Re-use of question forms was frequently mentioned, with just the latter words in the question being varied to assess different content.

4.1.5. Topicality

One subject-specific factor noted was topicality in government and politics. Given that these specifications are intended to reflect ongoing changes in the political landscape, use of topical questions could be considered to be appropriate, providing that the questions are not too obvious and these topical questions occur in an irregular way.

The participants' apparent underestimate of the lead time on exam paper production also make this a less useful predictive factor, as it is not always easy to think back to what was current at the point at which the questions would have been first written and selected for inclusion on a live paper.

4.2. Predictability of papers in this study

The measures of outcome space overlap obtained here may form the starting point for an objective measure of paper predictability. Taking outcome space overlap as a proxy measure for paper predictability, teachers were on average able to predict just under 40% of the knowledge and skills required to answer the psychology and history papers, and just under 50% for the government and politics papers.

The pre-reform government and politics papers had sometimes been criticised as overly predictable on social media. One of the government and politics papers was indeed the most predicted paper here, with over 60% predicted. This may be an indication as to where undesirable predictability lies on this scale.

History papers could on the surface appear to be quite predictable, with most high level topics appearing in some form on most papers, but in fact the number of different questions that can be asked and the different types of questions, with variety in command words, question stems and tariffs, may act to produce an appropriate level of predictability/uncertainty.

In some respects it may be a little surprising that the reformed psychology papers did not come out less predictable still, given their newness and the intention in the reform process to avoid predictable tests. However, a large part of their predictability lay in the skills they assessed, rather than the precise pairing of topic and skills.

These skills being predictable may not be a problem when the skills need to be applied in varied and unpredictable contexts.

This suggests that values of prediction overlap around 40% may indeed indicate good, or at least acceptable levels of predictability, while upwards of 60% may be worthy of attention. We can only speculate where an overly unpredictable test would fall, perhaps closer to 20%, although it may be that such a paper would vary by more qualitative features such as unexpected question forms, than just low overlap of live and predicted question outcome space.

4.3. Good vs bad predictability

It may in fact be that taking a single measure of paper predictability based on outcome space was not the best way to collect the ratings. The separation of skills and knowledge overlap which one of the experts in the exercise used might have been useful, as this may correspond well to the distinction between what we might call ‘good’ and ‘bad’ predictability.

The kind of skills assessed in an exam should be highly predictable so that candidates know what is expected of them. Asking a candidate to carry out a task or analysis in an exam which they would never have expected, and probably not been taught, may be a good test of general ability or intelligence, or adaptability at least, but would probably not be a fair or valid assessment of the subject construct.

A good assessment, from a predictability viewpoint, would assess known skills but the content area against which particular skills were assessed would be unpredictable - where there is some kind of constraint, or worse, a pattern, in matching between content and skills, we may have a concern over predictability.

The overlap of skills was largely a given in the current study. That is because the government and politics and history papers had a largely fixed structure, with set types of questions with fixed tariffs in each question slot on the paper. With largely fixed question stems, and the same assessment objectives for each question on each paper, broadly the same skills are assessed each time.

However, there may be minor differences in required skills due to the precise content specified. For these papers the rating of outcome space overlap, and therefore the final paper predictability metric, will mostly reflect the ‘bad’ predictability of topics coming up in predictable forms.

For the psychology paper, the unstructured nature of the papers meant that our experts were often quite focused on the skills element in their comments. By asking for an overall rating of outcome space overlap we combined skills/content and therefore ‘good’ and ’bad’ predictability. As discussed above, this makes our evaluation of psychology paper predictability a slightly different measure.

In a future repeat of this kind of approach, making clear the distinction between ‘good’ (skills) overlap and ‘bad’ (content) overlap and asking for separate ratings from experts would greatly help to clarify the extent to which paper predictability may be a concern.

4.4. Extensions to this research

Although this was quite a labour-intensive approach to obtaining a measure of predictability, there are modifications that could be used to produce more robust data on overlap for slightly less effort. Rather than holding a meeting to define a final predicted paper, individual teacher predictions of whole papers could be compared to the final paper, and the average and range of the prediction accuracy would give a more detailed measure of paper predictability.

The current study was designed to provide as much information as possible about the factors influencing predictions, rather than generating coherent final paper predictions. We could not easily evaluate the individual predictions because our participants were not directed to predict a final paper but to make as many predictions as they wished for each question slot on the paper. It would be a simple matter to adjust the instructions for each individual to predict a set of questions that would make up a single coherent paper.

Other approaches to predictability studies could include seeing how responses to one question are rewarded when they are entered as answers to a similar question. It is possible to imagine a situation where candidates may be taught in class to answer a particular question, but this precise question does not appear on the paper.

However, one of the questions on the paper could overlap with the prepared one to some extent, perhaps because the prepared question was less specific and so covered broader content, within which a live question falls. Seeing how answers that do not quite address the asked question are marked and treated by examiners would reveal whether questions/mark schemes are so broad and generic that preparing answers was a good tactic. Such papers/questions would then require reviewing and amending.

Pairs of questions with a substantial amount of overlap could be identified from different papers and responses to the ‘wrong’ question could be mixed into a set of responses to the ‘right’ question and the whole set remarked against the ‘right’ question mark scheme.

Statistical approaches could also be applied to determine where pre-prepared answers may have been taught in class. There is an existing literature on detecting cheating which covers different approaches in which unusual mark patterns within a centre or class can be diagnostic of cheating on a test. For more about this, see a review by He, Meadows and Black (2018).

Where questions on papers are predictable enough to make the preparation in class of prepared answers worthwhile, the marks may be more tightly clustered and probably higher than expected due to the over-performance of less able candidates when reproducing the prepared answer. Such an analysis of unusual mark patterns on longer answer questions could highlight for which tests teachers are successfully making predictions and preparing answers.

5. Conclusions

This study began with three research questions in mind. Through analysis of the reasons/rationales teachers described for the predictions they made, we have defined a large set of factors that teachers believe can be used to predict future questions. This was sometimes through a process of analysing past papers, sometimes through other wider sources of information such as the specification document itself, or supporting materials such as course textbooks. Following on from the teacher predictions, the selection of a set of final predictions making up complete question papers has allowed a method to estimate the accuracy of the predictions to be trialled, and thus to give a proxy measure for the predictability of the papers themselves.

Though only small scale, this has given comparative estimates of paper predictability for the papers in our study. Given some of the prior views on the papers, and the range of predictability scores we obtained, a start can be made at estimating where over- and under-predictable assessments may lie on this scale.

This work has provided a rich source of data for test developers and item writers to consider when trying to avoid overly-predictable tests. Awareness of the kind of factors that are used by classroom teachers to help make predictions should help to avoid overly predictable questions and papers. Awareness of these factors may help both when developing new qualifications, in terms of the content specification, paper design and questions on the sample assessments, and also when writing question papers during the lifetime of a specification.

Where possible, generating questions that do not follow the patterns defined by the framework of factors identified here will lead to less predictable tests. This, in turn would reduce the contribution of overly-predictable question papers to the sawtooth effect during the introduction of new qualification specifications, as described in Newton (2020).

The measures of paper predictability obtained here, as well as the suggestions on how to streamline and strengthen the approach, may form the first step in a new method to objectively measure paper predictability. However, this work cannot yet define conclusively where ‘good’ or ‘bad’ predictability lie on the measurement scale.

References

Baird, J.A, Caro, D.H. and Hopfenbeck, T.N. (2016). Student perceptions of predictability of examination requirements and relationship with outcomes in high-stakes tests in Ireland. *Irish Educational Studies*, *35(*4), 361-379, DOI: 10.1080/03323315.2016.1227719

Baird, J.A, Hopfenbeck, T.N., Elwood, J., Caro, D.H., and Ahmed, A. (2014b). *Predictability in the Irish Leaving Certificate*. Oxford University Centre for Educational Assessment Report. OUCEA/14/1. Retrieved from https://www.examinations.ie/about-us/Predictability-Overall-Report.pdf

Baird, J.A., Hopfenbeck, T. Ahmed, A., Elwood, J. Paget, C., and Usher, N. (2014a). *Predictability in the Irish Leaving Certificate Examination. Working Paper 1: Review of the Literature*. Oxford University Centre for Educational Assessment. Retrieved from https://www.examinations.ie/about-us/WP1-Review-of-Literature-and-Media-Analysis.pdf

Byrne, M., and Willis, P. (2004). Leaving certificate accounting: Measuring students' perceptions with the course experience questionnaire. *Irish Educational Studies*, *23*(1), 49–64, DOI: 10.1080/0332331040230107

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46. DOI: 10.1177/001316446002000104

Crisp, V., Sweiry, E., Ahmed, A., and Pollitt, A (2008). Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions, *Educational Research*, *50*(1), 95-115, DOI: 10.1080/00131880801920445

Daly, A., Baird, J., Chamberlain, S., and Meadows, M. (2012). Assessment reform: students' and teachers' responses to the introduction of stretch and challenge at A-level. *The Curriculum Journal*, *23*(2) 139–155, DOI: 10.1080/09585176.2012.678683

He, Q., Meadows, M., and Black, B. (2019). *Statistical techniques for studying anomaly in test results: a review of literature*. Ofqual report 18/6355/5. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/690007/Statistical_techniques_for_studying_anomaly_in_test_results-_a_review_of_literature.pdf

Johnson, M., Constantinou, F., and Crisp, V. (2017). How do question writers compose external examination questions? Question writing as a socio-cognitive process. *British Educational Research Journal*, *43(*4*)*, 700-719. DOI: 10.1002/berj.3281

Johnson, M. and Rushton, N. (2019). A culture of question writing: Professional examination question writers' practices. *Educational Research*, *61*(2), 197-213. DOI 10.1080/00131881.2019.1600378

Murphy, R., Stobart, G., Baird, J.A. and Winkley, J. (2012). *Investigating the Predictability of GCSE Examinations*. Pearson Internal Report. [Google Scholar]

Ofqual (2008). *Predictability Studies Report on GCSE and GCE Level Examinations*. Office of the Qualifications and Examinations Regulator. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605251/0808_Ofqual-Predictability_Studies_final_report.pdf

Newton, P. (2020). *What is the Sawtooth Effect?* Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/publications/what-is-the-sawtooth-effect

Pollitt, A. and Ahmed, A. (2008). *Outcome space control and assessment*. A paper presented at the 9th Annual Conference of the Association for Educational Assessment – Europe, Hissar, Bulgaria, November 2008. Retrieved from http://www.lifeinbits.org/camexam/htdocs/papers/OutComeSpaceAPAA.pdf

Appendix A - Structure of the papers

Here we review the structures of the papers included in the study, since these structures will influence how teachers make predictions, and the degree of freedom they have when predicting each question.

A.1 Specification 1 government and politics

Paper 1 and 2 both follow the same structure. Each has 12 questions, divided into 4 sections. These 4 sections are the same each year and appear in the same order on the papers. They mirror the 4 high-level content areas for the paper in the specification. Within each section are 3 questions with the same tariff structure for each section. A written source is provided at the start of each section, and the first 2 questions in the section relate to this source, with the final (highest tariff) question usually covering a different content area within the section's high-level content area. Candidates must answer the questions within any 2 of the 4 sections.

A.2 Specification 2 government and politics

The 2 papers here have slightly different structures, although both have 4 sections, which reflect the 4 main topic areas in the specification content for each paper. The 4 sections are the same each year, although the order of the sections varies on the papers.

Paper 1 is similar to the specification 1 papers in that there are 12 questions, with 3 questions in each of the 4 sections on the paper. The same tariff structure is used for each section. There are no source extracts on this paper. The order of the topic sections is different on each paper although this shuffling serves no purpose in terms of reducing predictability. Candidates much answer the questions in any 2 of the 4 sections.

Paper 2 is organised slightly differently, in that there are 2 sections with 3 questions each and 2 sections with a single large question each. The content areas are rotated around, so that sometimes they occur in the three-question sections, sometimes in the one essay question section. The 2 sections divided into 3 questions have a source extract at the start, and the first 2 questions refer to this source, with the third usually covering different content within the main topic area. The 2 other sections comprise a long essay question with no source material. Candidates must answer the questions in 2 of the sections, one of the 2 three-question sections, and one of the 2 long essay question sections.

A.3 Specification 1 history

This paper has 3 main sections. Part 1 comprises 2 sections that have the same structure of questions stems and tariffs as each other and candidates must choose between 2 sections on different historical areas. Each of these 2 sections is made up of 8 questions with a mirrored structure. Within each optional route, the first 2 compulsory questions are on one content area, with the first a source-based question followed by a related question. Then there are 2 optional sets of 3 questions on different topic areas. These again have a fixed tariff structure and the questions follow the same pattern in terms of the skills and cognitive operations required of candidates. Within each optional route candidates therefore answer questions on 2 of the 3 high-level content areas.

Part 2 is the depth study section which is compulsory. This starts with 3 compulsory source-based questions. This is then followed by 2 optional sets of 3 questions which the candidate has to choose between. Apart from additional spelling, punctuation and grammar marks in the Part 2 sections, these groups of 3 questions follow the same pattern in Part 1 and Part 2. Again, candidates answer the questions on 2 out of the 2 presented sections. Because of the optionality on this paper there are more questions overall than across the 2 equivalent specification 2 papers.

Although the paper structure is fixed in terms of the 3 main sections (two in Part 1, and Part 2), within each section the specification content is rotated around the different groups of questions each year. There is a reasonable amount of optionality within each section in these papers, although candidates will often not be able to choose between the 2 sections in Part 1 as they will often only have been taught one of the areas.

A.4 Specification 2 history

This specification splits the content across 2 papers, with Paper 1 covering content which is similar to specification 1 Part 1 and Paper 2 covers similar content to the specification 1 Part 2 depth study, although the content does differ slightly between the exam boards.

Both papers have a fixed question structure every year, with the same tariffs and the same types of question (generally the same question stems) in each position on the papers, but the content areas are rotated around the questions, and several questions draw on content from more than one main content area.

Paper 1 has a source question with a follow-up second question. The third question includes 2 different sources. Next, there is a choice of one from 2 questions followed by a final compulsory question in which the candidate has to write about 3 of the 4 topic areas presented.

Paper 2 starts with a set of 4 compulsory questions, the first of which refers to a source. This is then followed by 2 questions where the candidate has to choose to answer one, followed by another 2 questions to choose between, both which include 2 hints to refer to in the answer.

A.5 Specification 1 psychology

This paper is divided into 4 sections, each assessing a main sub-division of the specification content for the paper, covering 4 areas in psychological research. Although each section has an equal number of marks (24), there are a variable number of questions in each section, which are distributed across the content that each section addresses. Over the entire paper there are certain question types that must come up, including extended essays, and questions assessing skills and knowledge in research methods, practical research and mathematical skills. These questions can occur in any section. There is no optionality – every question must be attempted.

A.6 Specification 2 psychology

This paper has 3 sections, each carrying equal weight of marks. Section A covers a series of 5 topic areas, each of which contains 2 themes. Each theme comprises 2 key research papers. This section has a fixed total tariff but there is flexibility in the number and type of questions asked (while maintaining a constant balance across assessment objectives). Not every research paper is assessed, and in fact not every theme is assessed on each paper although each of the 5 topic areas will be. Question tariffs and types vary although these are mostly lower-tariff questions.

Sections B and C are semi-structured, with similar, but not identical types of questions within each section on each paper, and with a little bit of variation in the number of questions and their tariffs. Section B assesses the content area of 'areas, perspectives and debates' with a focus on the similarities, differences and inter-relationships between different areas of psychological research, while Section C is a practical applications section, with a source used to introduce an issue for the candidate to analyse. There is no optionality on this paper.

Appendix B – Instructions to teacher predictors

Predictability Framework

Introduction

Your task is to consider the specification document and the set of past (or practice) papers we have provided and to think about questions that may appear in the next question paper in the series in Summer 2017.

You probably already write questions for your students to work on in class or homework exercises. The particular focus here is on those questions that you might expect to appear on the Summer 2017 question paper.

On the following pages, we have tried to design a document that reflects the way the papers are structured, and to ask for a level of detail in your predictions which is reasonable for you to make. There are sections for you to complete which involves taking into consideration the topics and/or specific questions for the different sections of the paper that you think will come up in Summer 2017, and to record your reasoning around the choices and suggestions you make.

The recording of your thought processes when suggesting questions/topics is vital to the success of this project. We are looking for a deeper level of analysis than just ‘this usually comes up’ ☺. We would like you to consider factors such as:

- Alignment of wording in the specification with questions
- Curriculum coverage of assessments
- Granularity of questions
- Number of questions
- Question tariffs/item types
- Paper format, i.e. sections, optional routes
- Use of sources/case studies in question papers
- Specificity of sources/case studies in questions that hang off it.
- Longevity of specification and cycling of topics or questions
- Revision guides/textbooks

This list is not in any way exhaustive, and there may be other factors that inform your thinking. So, please feel free to indicate whatever factors influence you choice.

Notes

1. When writing a question, we are not worried about the exact wording of the question, so please do not agonise over this. We are interested in the topic you have chosen, and the way you have chosen to assess it.
2. This is not a whole-paper design exercise. For example, there is no need to worry about balancing Assessment Objectives weightings when suggesting questions.

3. If you can think of more than one question that is likely to come up for a question slot on the paper, please make additional suggestions and explanations for each question – don’t be limited to only one, just expand the table cell for your suggestions.
4. Use the ‘probability’ column in the tables to indicate how probable you think it is that your suggested question/topic will come up. Please use a scale of 0-100%, where 0% indicates ‘This will not come up’ and 100% indicates ‘This will definitely come up’. If you make more than one suggestion for a question please give one of them a higher probability than the others – this is your top choice.

Appendix C – Instructions to subject expert question overlap raters

Summary of task

We want to gain an understanding of how much of the actual paper sat by students in summer 2017 was correctly predicted in the attached “predicted papers”. [Alongside this document you should have received **four** predicted papers in separate pdf documents.]

You may want to think of the overall task this way – by evaluating individual questions for their overlap, we hope to gain a picture of how much of the actual summer paper a candidate would be able to answer if they knew absolutely nothing *except* how to produce full answers to all of the questions on the predicted paper.

We would like you to consider each pair of **actual** and **predicted** papers. In this work package there are 4 pairs:

- Actual [specification 1] Paper – Predicted [specification 1] Paper A
- Actual [specification 1] Paper – Predicted [specification 1] Paper B
- Actual [specification 2] Paper – Predicted [specification 2] Paper A
- Actual [specification 2] Paper – Predicted [specification 2] Paper B

To help with this task, this document lists the actual questions from each paper twice, with each set to be compared independently to just one of the predicted papers as above.

How to complete this document

We would like you to work through each question listed in this document and judge how similar it is to the closest matching question on the indicated predicted paper, in terms of the overlap of the ‘outcome space’ of the actual question (see below). We would also like you to detail your thinking and reasons behind the rating you give.

The main body of this document consists of tables with 4 columns:

1. The actual question of concern

2. A column to write the number of the best-matching predicted question

3. A column to write your similarity/overlap rating from 0 to 10

4. A column to describe your thinking behind this rating. Please give as much detail as necessary to justify your rating.

Making the similarity judgement

We would like you to think about something we call the ‘outcome space’ of a question. By outcome space we mean the range of possible answers that a question is designed to elicit. The Appendix at the end of this document contains a summary adapted from a conference presentation describing the use of this approach in writing questions, which gives a good feel of what the outcome space represents.

The idea here is to think about the outcome space of the **actual** question, and consider how much of this outcome space the closest **predicted** question overlaps.

As a guide:

0 indicates no overlap at all, no part of the actual question would be answerable to a candidate only knowing the answers to the predicted questions.

5 indicates 50% overlap of the outcome space of the actual question. This may not equate exactly to 50% of the marks, but would indicate the candidate trained on the predicted questions would have about half of the knowledge and skills required for the actual question.

10 indicates total overlap of the outcome space of the actual question. This may be through the questions being effectively identical – note that is important that the *question wording does not need to be the same*. However this rating is also possible if the outcome space of the actual question is a subset of the outcome space of the predicted question, provided the outcome space of the predicted question is at the same level of detail.

So for each actual question in this document, we want you to pick the predicted question which encompasses the **largest proportion** of the outcome space of the actual question.

Important points to note:

- It is fine to use the same predicted question to match against more than one actual question if this explains the most outcome space on the actual paper.
- You may find the best matching predicted question anywhere on the predicted paper, although it is likely to be in a similar section due to the way the papers are structured.
- There is no need for the best-matching predicted question to have the same tariff.
- Specific question wording does not matter – it is the outcome space these words are defining which matter – it is possible that quite different wordings can define the same outcome space.
- Outcome space is a combination of both knowledge and skills.
- There is no definite right or wrong overlap rating – different individuals may apply the 0-10 scale slightly differently – this is fine, the important thing is to be consistent across your own judgements.

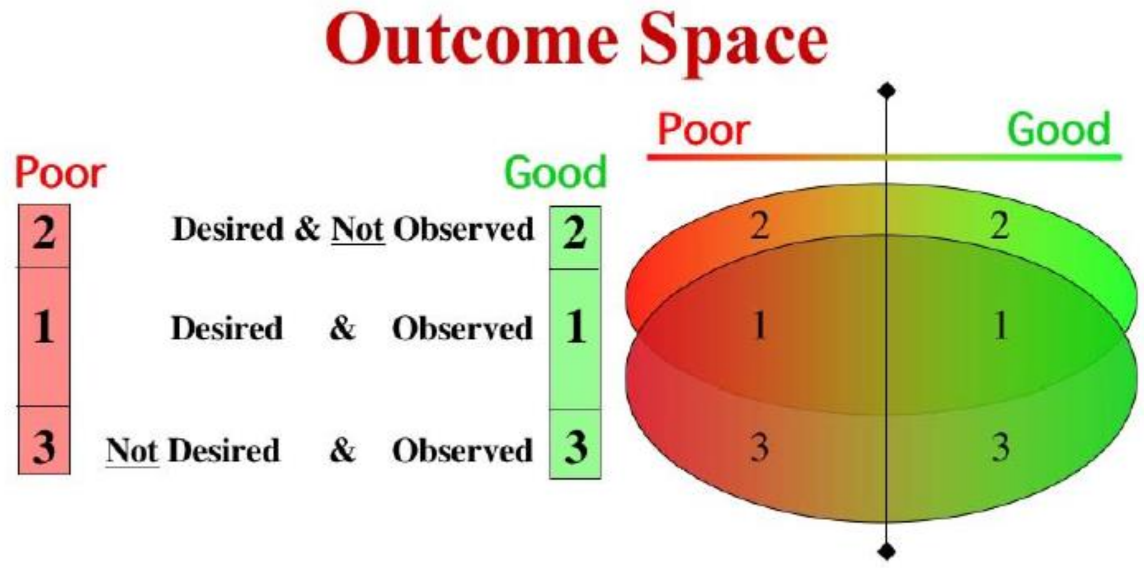
So, to restate the overall task– if a candidate knew absolutely nothing except how to produce full answers to all of the questions on the predicted paper, what is the largest proportion of the actual paper they would be able to answer.

Note that we have not included mark schemes for the actual questions. Given that the predicted questions do not have mark schemes, we want you to imagine the outcome space of the questions using the question alone.

Appendix D – Outcome space control and assessment

Pollitt and Ahmed devised Outcome Space Control and Assessment (OSCA) as a framework for writing assessment items. By considering the kinds of response that a question elicits, and the kinds of response that the mark scheme rewards, it is possible to ensure valid and effective questioning to assess a trait.

Outcome Space simply refers to the evidence – the responses examiners hope to get as evidence of understanding of the trait, and the responses the students produce. The concept of Outcome Space (OS) comes from work by Marton and Saljo (1976) in which they used the term to describe the range of responses students produced when asked questions on an academic article. They used OS to describe the qualitative differences in responses. Pollitt and Ahmed use the term in a more general way to refer to exam question responses. For any exam question there will be a range from poor to good responses that students will produce. The diagram below shows how the OS can be divided up:



The areas *Poor 1* and *Good 1* represent the answers the question is meant to elicit and does. Note that it is very important to consider all of the poor answers as well as the good ones, if the question is to succeed in its aim of validly discriminating between poor and good students. For valid assessment we would like these zones to be as large as possible, as they indicate students behaving as the examiners intended.

Poor 2 and *Good 2* represent responses that the examiners expected to see but that did not in fact occur; this would include any alternative good but obscure answers, as well as anticipated errors that didn't happen. Some of this space – especially *Poor 2* – is unavoidable, but examiners should at least pause to think why no students came up with errors that the examiners expected them to make. Perhaps the question wording allowed students to avoid these errors?

Poor 3 and *Good 3* are more problematic in terms of validity, as they represent outcomes that were not anticipated by the examiners and cannot, by definition, be included in an initial mark scheme. Any frequently occurring answers in the *Poor 3* zone may indicate a way in which the question could plausibly be misunderstood, an ambiguity or an unfair distraction, and shows that the examiners had lost control of the students' thinking processes. Any response that has to be classified as *Good 3*, an unanticipated but correct response, is more obviously an indication that the examiners had lost control of the question.

As an example, the concept of outcome space is used to describe how exam tasks should be written in a systematic way in order to maximize validity.

The process of question writing must begin with an understanding of the trait examiners are trying to measure, and what it means to be 'good' or 'poor' on that trait. With an idea of a task in mind, they will then be able to decide on what evidence they would like to see that will help them to discriminate between good and poor performances on the task. This evidence, i.e. performance on a task, is called the Desired Outcome Space.

Next they should consider how they intend to make inferences about the students' competence from their performances, which means they should draft the mark scheme. The crucial issue is how to infer which students have greater competence on this trait – the mark scheme, if it is based on a careful consideration of the evidence, the desired Outcome Space, should facilitate this. Only then do the examiners consider how to elicit the performances they want by working on the precise wording of the question.

There then follows a process of iteration around question, desired outcome space and mark scheme, until the examiners arrive at a question and mark scheme that will elicit the desired outcome space and evaluate it appropriately.

Following the systematic procedure outlined in OSCA is the best way to ensure valid assessment. This is achieved by eliciting evidence of the right kinds of mental behaviour - *the things we want them to show us they can do* – and by evaluating the resulting performances in order to make valid inferences about competence.

Adapted from Pollitt and Ahmed (2008), "Outcome space control and assessment." A paper for the 9th Annual Conference of the Association for Educational Assessment – Europe. Accessed Jul 2017 from http://www.lifeinbits.org/camexam/htdocs/papers/OutComeSpaceAPAA.pdf

Appendix E – Comparison of prediction template coding approaches - either coder vs both coders

This section includes figures showing the similarity between factor coding that is restricted to where both coders agreed on the factor (both raters only), or factor coding including where only one of the two raters coded the factor (either rater).

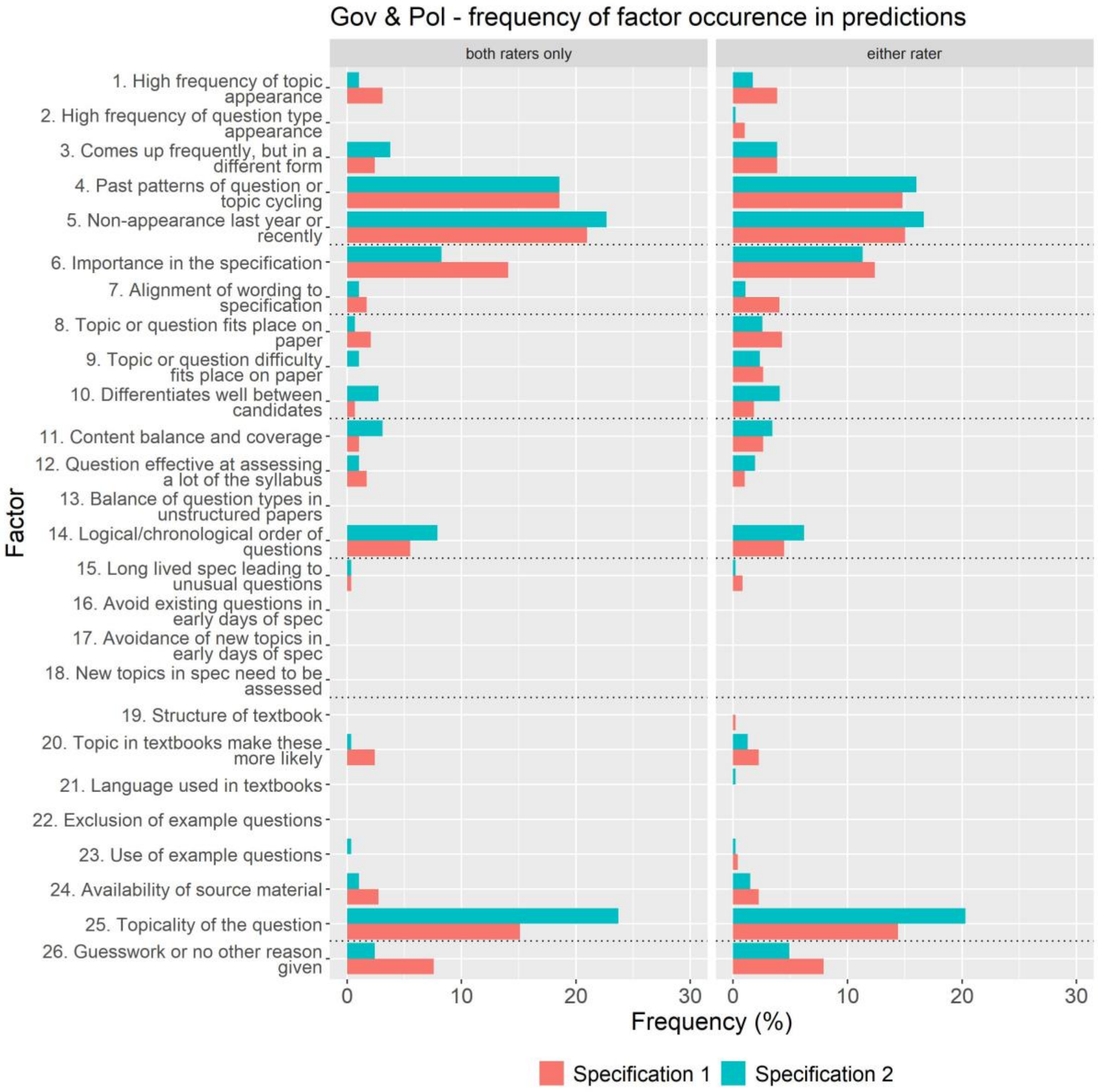


Figure E1: *Frequency of occurrence of each different factor in individual predictions made by teachers for government and politics. Specifications are distinguished by colour with the two coding approaches in the two panels.*

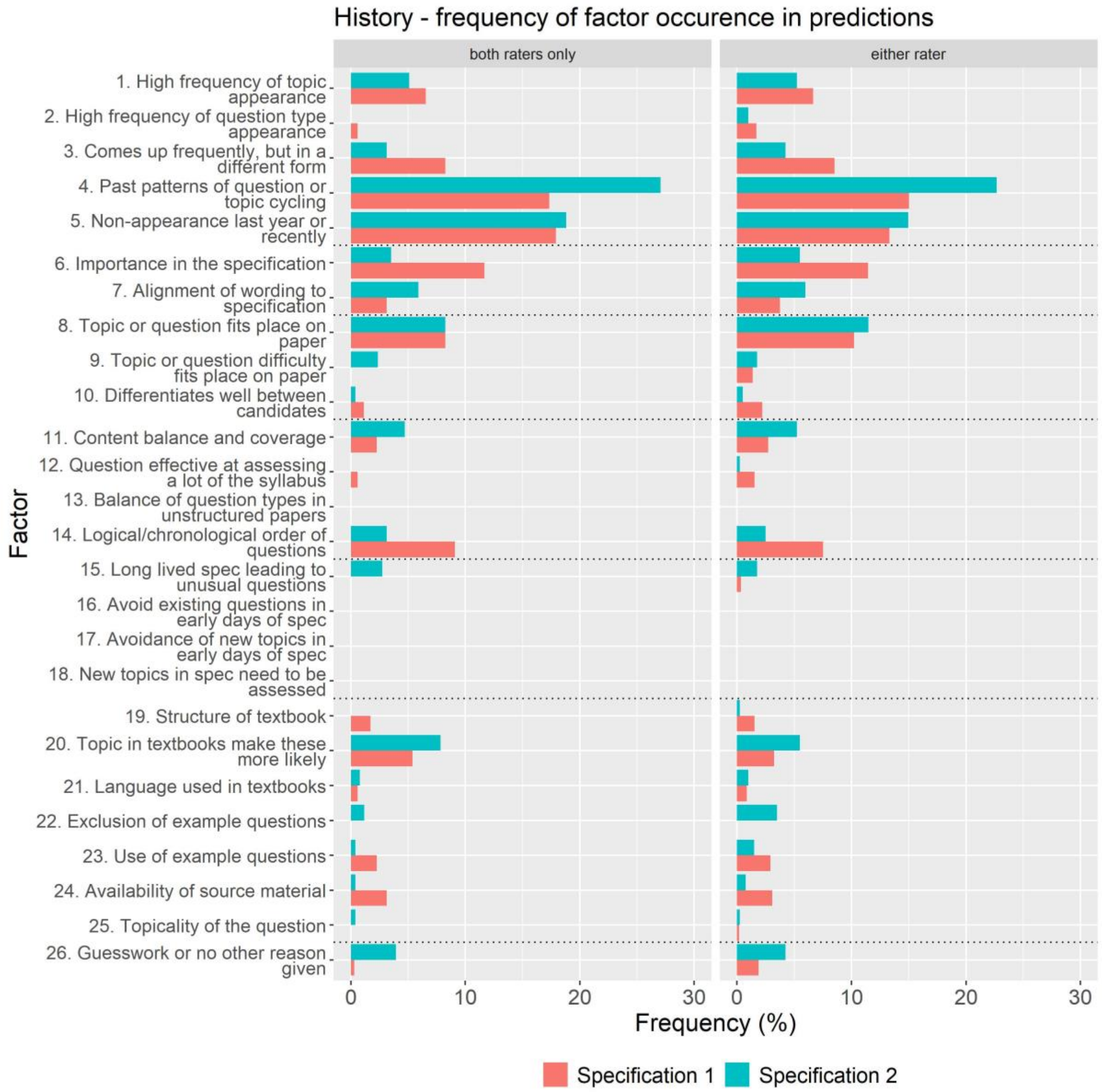


Figure E2: *Frequency of occurrence of each different factor in individual predictions made by teachers for history. Specifications are distinguished by colour with the two coding approaches in the two panels.*

Across all government and politics and history the factor frequencies are spread more widely when either rater coding is used, since that coding includes more in less frequently-used categories. The relative proportions between the most frequently coded factors do not change significantly between coding approaches. The main body reports only the either rater coding.

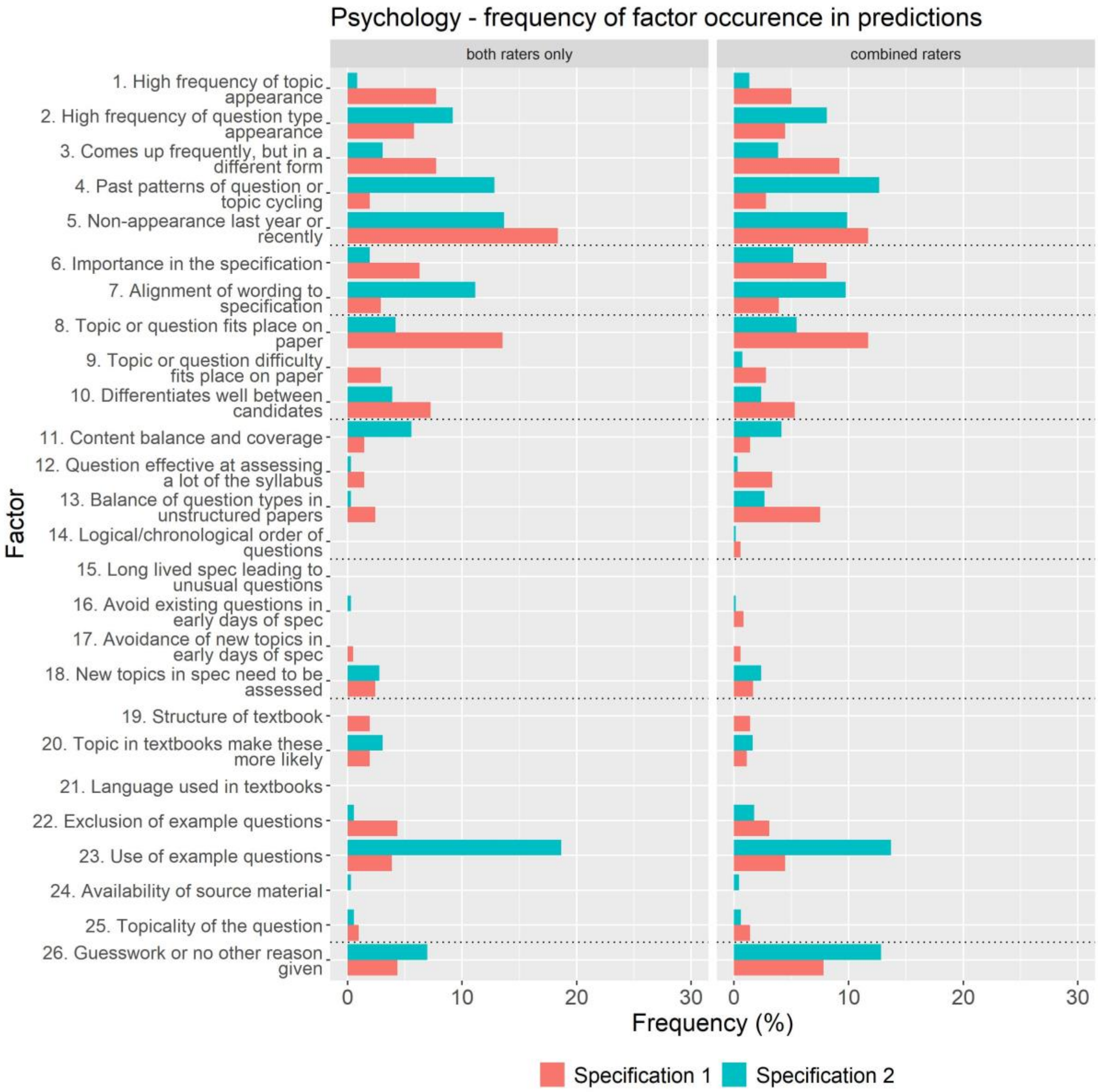


Figure E3: *Frequency of occurrence of each different factor in individual predictions made by teachers for psychology. Specifications are distinguished by colour with the two coding approaches in the two panels.*

The psychology coding was slightly different, as it consisted of counts of occurrences of factors within each paper section, rather than the binary yes/no per question slot for the other subjects. The comparison in Figure E3 is therefore between the taking the lowest of the two counts (both raters only – we assume they are agreeing on all the coding by taking the lower number) or taking the higher of the two counts (combined raters) which we assume represents cases where either rater would have coded the factor. The main body reports only the combined rater coding.

Appendix F – Teacher predictor survey responses

All 55 teachers taking part in phase 1 completed a survey after returning their prediction document. We were interested to know whether the predictability of specifications was a major concern for teachers, and also whether they tried to predict topics and questions that might appear on papers in the future, and how they used these predictions.

Following questions relating to their teaching experience (reported in the main body of this report) they answered a series of fixed-response questions. When asked “Why do you teach your current specification?”, Figure F1 shows that responses were similar across subjects, with content and being good for the students as prime motivations in teaching the chosen specification.

One significant difference between subjects was in the predictability of the specifications, with this being a stronger motivation for the government and politics teachers than for the other subject teachers. Other reasons given by a small number of teachers in a free text field were being an examiner for that specification, having inherited the specification when joining the department and comments giving more detail around why it worked well for students.

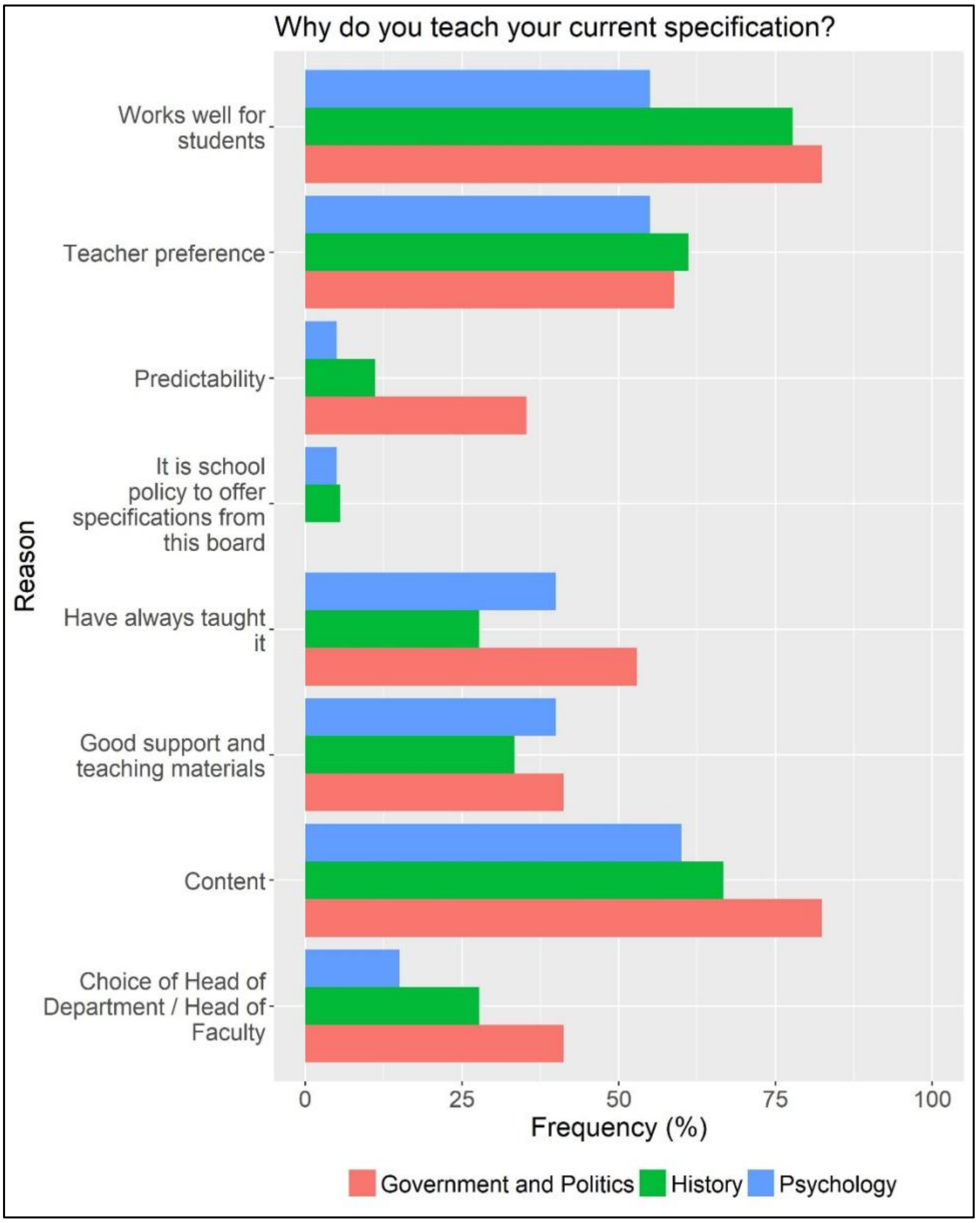


Figure F1: *Responses to the question "Why do you teach your current specification?" split by subject. (N=55).*

We then asked whether the teachers felt that their specification was more or less predictable than other specifications in the subject (Figure F2). Many respondents did not know, presumably due to lack of familiarity with other specifications, and where an opinion was expressed generally there was thought to be little difference in predictability across specifications. No psychology teacher thought their specification was more predictable than other specifications.

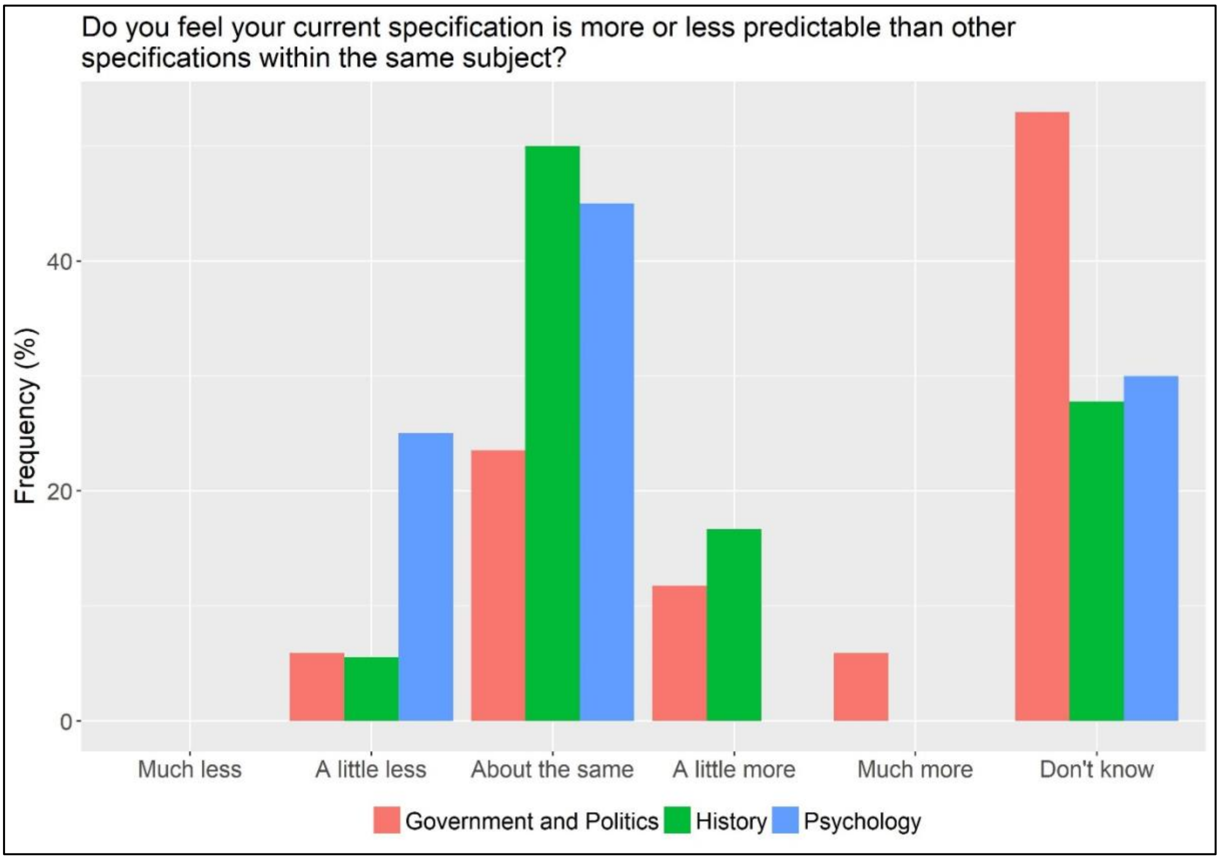


Figure F2: *Responses to the question "Do you feel your current specification is more or less predictable than other specifications within the same subject?" split by subject. (N=55).*

When asked whether they predicted general topic areas that they expected to appear on future papers, around half of the teachers reported that they did so (see Figure F3). History teachers were more likely to report doing so, and psychology teachers less likely to do so.

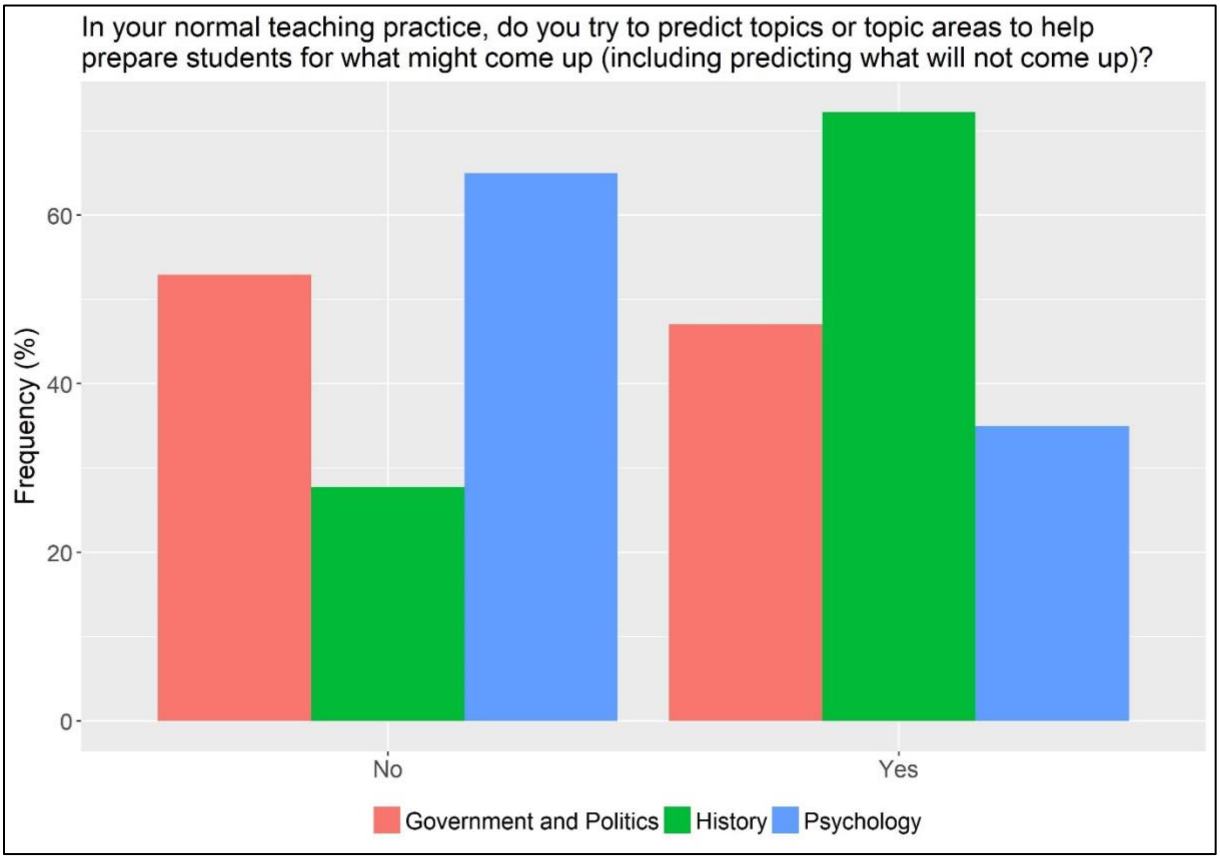


Figure F3: *Responses to the question "In your normal teaching practice, do you try to predict topics or topic areas to help prepare students for what might come up (including predicting what will not come up)?" split by subject. (N=55).*

The follow-up question for those who said they did predict topics asked about the prediction of actual questions that might come up. Of those teachers who predicted topics, the vast majority also tried to predict specific questions (see Figure F4). This was particularly true for government and politics and history, suggesting that some psychology teachers may be more focused on general topics than actual questions.

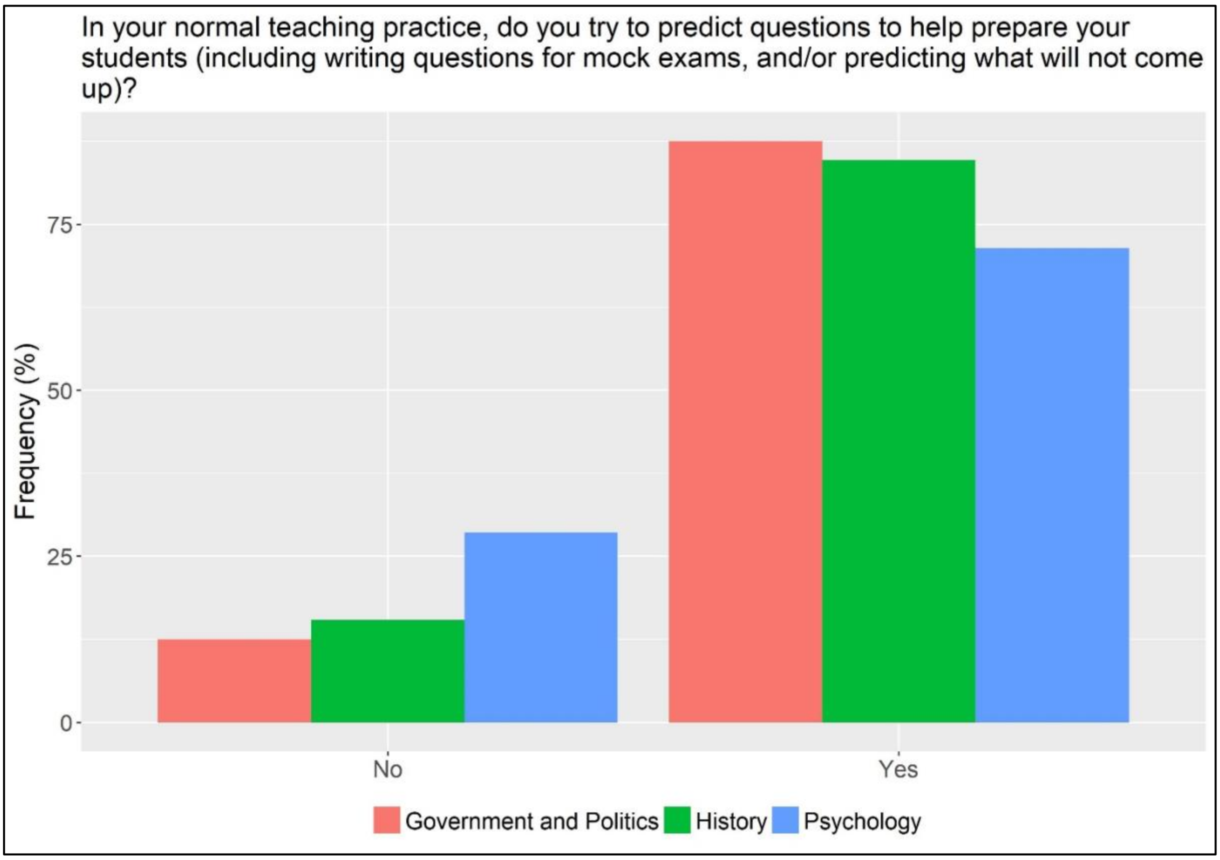


Figure F4: *Responses to the question "In your normal teaching practice, do you try to predict questions to help prepare students for what might come up (including predicting what will not come up)?" split by subject. This question was only asked of those teachers who said they predicted topics and so the percentages above relate to that subgroup. (N=55).*

Teachers who reported predicting topics were asked how they used the predictions they made (see Figure F5). The main uses were in preparing tests and also class exercises/homework. This latter use was much more frequently reported by government and politics teachers, and least frequently reported for psychology teachers. Psychology teachers were marginally more likely to use these predictions for deciding which topics to focus on in class rather than just exercises or homework. This focus on topics is consistent with the lower likelihood of psychology teachers predicting actual questions. Around half of history and government and politics teachers did use these predicted questions as prepared answers for exams. The only other notable difference was that government and politics teachers were less likely to share their predictions with others.

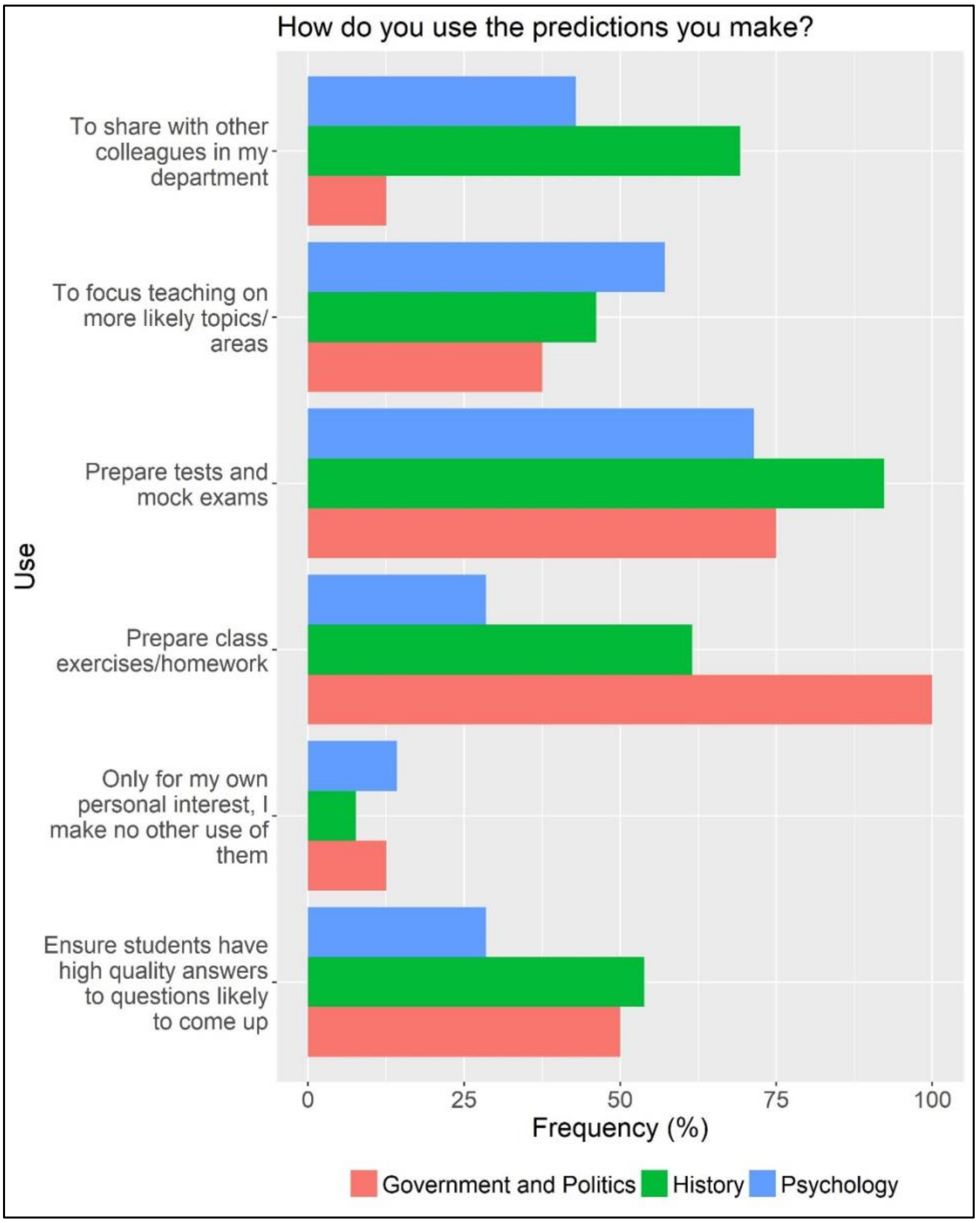


Figure F5: *Responses to the question “How do you use the predictions you make?” split by subject. This question was only asked of those teachers who said they predicted topics and so the percentages above relate to that subgroup. (N=28).*

Other uses suggested by teachers were for student revision purposes, while a couple of teachers said they used predictions to illustrate how hard it is to make them, and that students really needed to know the whole syllabus.

When those teachers who predict topics were asked how good their predictions have been in the past, most predictions were felt to be quite accurate (see Figure F6). Only a couple of psychology teachers felt their predictions were generally very accurate. Overall psychology and history teachers reported higher accuracy in their predictions, and government and politics lower accuracy.

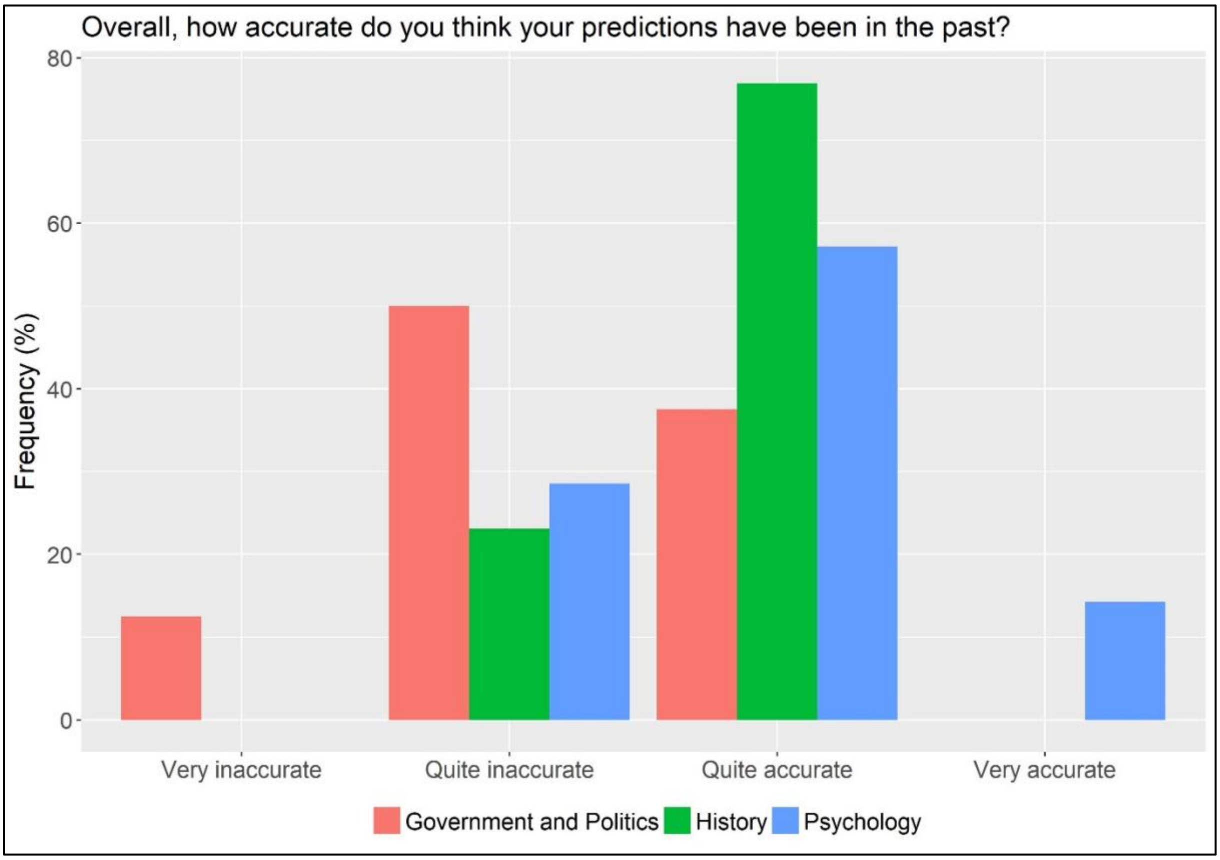


Figure F6: *Responses to the question "Overall, how accurate do you think your predictions have been in the past?" split by subject. This question was only asked of those teachers who said they predicted topics and so the percentages above relate to that subgroup. (N=28).*

The same subgroup of teachers making predictions were asked what impact they thought their predictions had on their students' grades in the exams (see Figure F7). Small to moderate impacts were generally reported, and there were no major differences between teachers in different subjects. All government and politics teachers thought that their predictions had some impact at least, while a few history and psychology teachers thought their predictions had no effect.

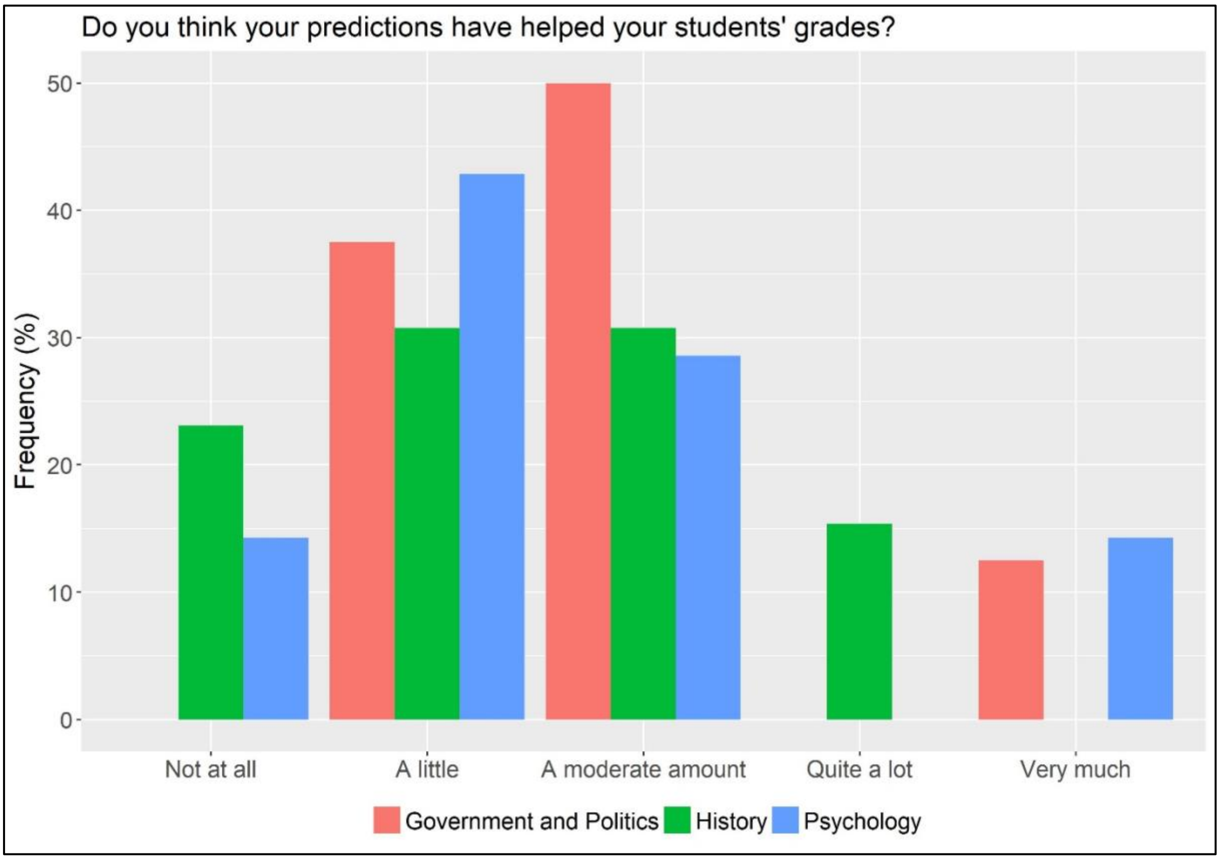


Figure F7: *Responses to the question "Do you think your predictions have helped your students' grades?" split by subject. This question was only asked of those teachers who said they predicted topics and so the percentages above relate to that subgroup. (N=28).*

Similarly the same teachers thought that the predictability of the specification as a whole did not have a large impact on students' grades (see Figure F8). Government and politics teachers reported the biggest impact from specification predictability, with psychology and history teachers less.

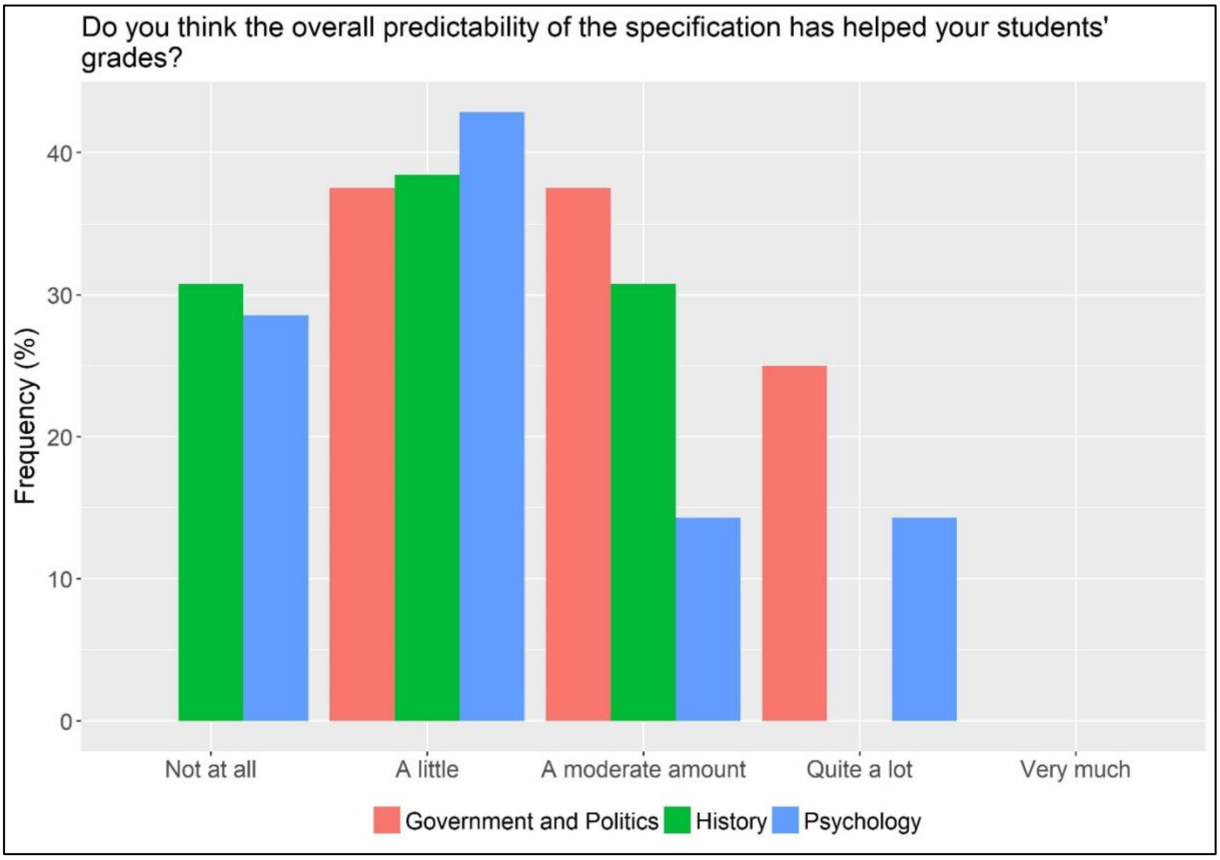


Figure F8: *Responses to the question "Do you think the overall predictability of the specification has helped your students' grades?" split by subject. This question was only asked of those teachers who said they predicted topics and so the percentages above relate to that subgroup. (N=28).*

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

December 2020 **Ofqual/20/6714/1**